AD-A285 932

The Pennsylvania State University
APPLIED RESEARCH LABORATORY
P.O. Box 30
State College, PA 16804

A WAVELET MODEL FOR VOCALIC SPEECH COARTICULATION

by

R. C. Lange

DTIC
ELECTE
NOV 04 1994
S B

Technical Report No. TR 94-13
October 1994

Supported by:
Space and Naval Warfare Systems Command

L.R. Hettche, Director
Applied Research Laboratory

Approved for public release; distribution unlimited

94-34315
94 11 3 08 1

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1994	

4. TITLE AND SUBTITLE

A WAVELET MODEL FOR VOCALIC SPEECH COARTICULATION

5. FUNDING NUMBERS

N00039-92-C-0100

6. AUTHOR(S)

R. C. Lange

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Applied Research Laboratory
The Pennsylvania State University
P. O. Box 30
State College, PA 16804

8. PERFORMING ORGANIZATION REPORT NUMBER

TR# 94-13

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Space and Naval Warfare Systems Command
2451 Crystal Drive
Arlington, VA 22245-5200

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

A known aspect of human speech is that a vowel produced in isolation (for example, "ee") is acoustically different from a production of the same vowel in the company of two consonants ("deed"). This phenomenon, natural to the speech of any language, is known as consonant-vowel-consonant coarticulation. The effect of coarticulation results when a speech segment ("d") dynamically influences the articulation of an adjacent segment ("ee" within "deed").

A recent development in the theory of wavelet signal processing is wavelet *system* characterization. In wavelet system theory, the wavelet transform is used to describe the time-frequency behavior of a transmission channel, by virtue of its ability to describe the time-frequency content of the system's input and output signals.

The present research proposes a wavelet-system model for speech coarticulation; wherein, the system is the process of transformation from a *control* speech state (input) to an *effected* speech state (output). Specifically, a vowel produced in isolation is transformed into an effected version of the same vowel produced in consonant-vowel-consonant, via the "coarticulation channel". Quantitatively, the channel is determined by the wavelet transform of the effected vowel's signal, using the control vowel's signal as the mother wavelet.

A practical experiment is conducted to evaluate the coarticulation channel using samples of real speech. The results show that the model is capable of depicting coarticulation effects associated with certain vowel-consonant combinations. They suggest that elements of the vowel's acoustic composition are continuously present, in a modified form, throughout the consonant-vowel transition. For other phonetic combinations, however, the model does not respond to instances of segmental transition in a characteristic way.

The conclusions drawn from the study are that the wavelet techniques employed here are effective tools for the general analysis of speech sounds, and can provide, in certain cases, a moderate enhancement over classical spectrographic methods. Similarly, the proposed coarticulation model does not reveal any specific acoustic/phonetic invariances in association with segmental coarticulation. It does, however, lay the groundwork for new approaches to analyzing the acoustic contingent of coarticulation within a systematic, potentially amendable, framework.

14. SUBJECT TERMS

wavelet, vocalic speech coarticulation, vowel, consonant, analysis of speech sounds

15. NUMBER OF PAGES

174

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

ABSTRACT

A known aspect of human speech is that a vowel produced in isolation (for example, "ee") is acoustically different from a production of the same vowel in the company of two consonants ("deed"). This phenomenon, natural to the speech of any language, is known as consonant-vowel-consonant coarticulation. The effect of coarticulation results when a speech segment ("d") dynamically influences the articulation of an adjacent segment ("ee" within "deed").

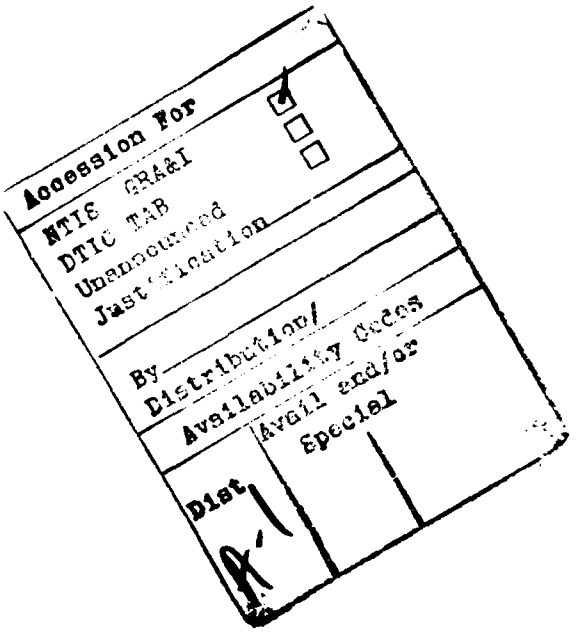
A recent development in the theory of wavelet signal processing is wavelet *system* characterization. In wavelet system theory, the wavelet transform is used to describe the time-frequency behavior of a transmission channel, by virtue of its ability to describe the time-frequency content of the system's input and output signals.

The present research proposes a wavelet-system model for speech coarticulation; wherein, the system is the process of transformation from a *control* speech state (input) to an *effected* speech state (output). Specifically, a vowel produced in isolation is transformed into an effected version of the same vowel produced in consonant-vowel-consonant, via the "coarticulation channel". Quantitatively, the channel is determined by the wavelet transform of the effected vowel's signal, using the control vowel's signal as the mother wavelet.

A practical experiment is conducted to evaluate the coarticulation channel using samples of real speech. The results show that the model is capable of depicting

coarticulation effects as ed with certain vowel-consonant combinations. They suggest that elements of the vowel's acoustic composition are continuously present, in a modified form, throughout the consonant-vowel transition. For other phonetic combinations, however, the model does not respond to instances of segmental transition in a characteristic way.

The conclusions drawn from the study are that the wavelet techniques employed here are effective tools for the general analysis of speech sounds, and can provide, in certain cases, a moderate enhancement over ... pectrographic methods. Similarly, the proposed coarticulation model does not reveal any specific acoustic/phonetic invariances in association with segmental coarticulation. It does, however, lay the groundwork for new approaches to analyzing the acoustic contingent of coarticulation within a systematic, potentially amendable, framework.



CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

Chapter 1

INTRODUCTION

1.1 The Use of Wavelets for Speech Analysis and Modeling

Many new developments in the theory and application of time-frequency signal analysis have been realized in recent years. Among them is the application of wavelet theory as a tool for characterizing and detecting time-varying signals. The wavelet transform generates a three-dimensional representation of a signal, using parameters which indicate the relative signal amplitude at various time-locations and various scale values.

Another (very recent) development in this area is the wavelet theory of *system* characterization. A system refers to any signal-transmission channel that is subject to an input excitation and results in a specific output response. The output signal depends on the particular attributes of the channel, with respect to the content of the input signal. In wavelet system theory, the wavelet transform can be used to describe the time-frequency behavior of a transmission channel by virtue of its ability to describe the time-frequency content of the system's input and output signals.

For the purposes of signal analysis, the wavelet transform is particularly well-suited to speech. As already indicated, it provides a signal representation which is a continuous function of time, thus capable of resolving the transient aspects of consonantal speech. In addition, the limits of time-resolution (and the corresponding limits of

frequency-resolution) inherent to the wavelet transform are linearly varying. This means that high-frequency components in a signal are resolved sharply in time, yet poorly in frequency. Conversely, low-frequency components are resolved sharply in frequency, yet poorly in time. This variable resolution is complementary to the distribution of time-frequency energy in most speech sounds. It should be noted that a specific class of wavelet transforms, the Morlet, is mathematically equivalent to the short-time Fourier transform, if the latter is formulated under windows of variable bandwidth.

Finally, the signal-model employed by the wavelet transform is capable of evaluating dips in the frequency spectrum as precisely as it evaluates spectral peaks. In this sense it can, again, be likened to the short-time Fourier transform. The distinction, however, represents a departure from an auto-regressive method (such as linear predicative coding) which relies on the location of a finite number of well-defined spectral peaks. The distinction suggests the usability of the wavelet transform for analyzing the classes of nasal speech sounds.

Yet, the number and scope of practical investigations into utilizing the wavelet transform for the analysis of human speech has been quite limited. For an example, see Kronland-Martinet et al. 1987. In general, these investigations indicate some distinct advantages afforded through this application. On the other hand, *no* investigations have been made into applying wavelet *system* theory to speech (for instance, as a means of describing human speech production). This research proposes a speech model which is an application of this theory. The model describes the time-frequency behavior of an effect common in human speech production, consonant-vowel-consonant coarticulation.

1.2 A Speech Model Based on Wavelet System Theory

The model proposed in this research relies on prior developments in wavelet system theory. In particular, the work of Young (1993, chapter 5) can be used as a basis for modeling the vocal tract as a time-varying transmission channel, subject to the time-varying excitation of the glottal source. The present model, however, extends these results and provides a means for describing the generation of speech production *effects*.

Here, a system is considered as the process of transformation from one speech state to another. The input to the system is one type of utterance. The output of the system is a variation on that same utterance. By virtue of the system channel which performs the operation, the speech "effect" is depicted as the transformation from input to output. The channel therefore constitutes a "speech-effect transform." This model is illustrated in Figure 1.1 below:

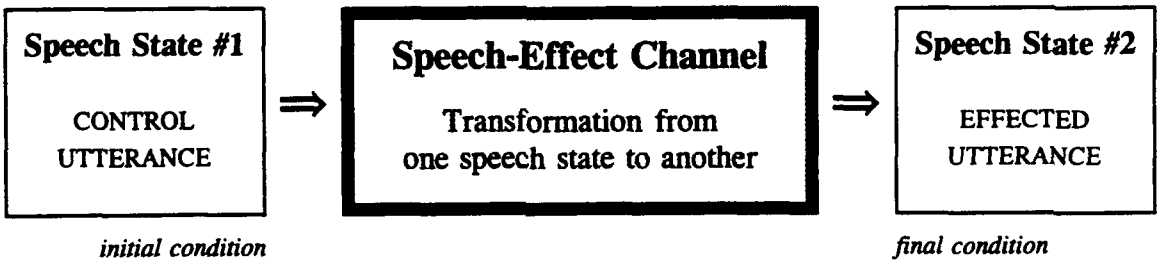


Figure 1.1 **The Speech-Effect Transfer System**

The modeled speech effect might be any speech condition known to influence speech production, such as voice quality, the influence of a particular phonetic context, or the differences between speakers. Because the system (input, channel, output) is formulated in wavelet-transform terms, the model is capable of describing speech effects in a time-varying fashion (i.e., in the form of a time-frequency distribution).

The principal structure of the speech-effect model assumes the form of a comparison between two utterances. An utterance is not characterized absolutely by the model; rather, one utterance is characterized *relative* to a another. Such a characterization provides a direct description of the contrasting effect. Whereas a more traditional method of characterizing a speech effect might require two stages of analysis (one for the effected utterance and another for the control utterance); the present model shows the difference or "transition" (from control to effected) within a unified description.

1.3 Significance of the Coarticulation Problem

The classical means of modeling the human speech production mechanism identifies a limited number of independent parameters whose interaction elicits the wide variety of possible speech sounds. In such a model, each parameter or parameter-group often represents the behavior of a specific articulatory structure. (For example, the larynx may be represented by one parameter group and the vocal tract by another.) The parameter representation might be based on a structural model or on a signal model. In either case, however, the model is typically resonant, wherein, a set of *static* parameter

values are specified for each phonemic sound "segment". A compound utterance, composed of a series of many such segments, is thus modeled by a series of successive parameter evaluations. Each parameter evaluation is discrete and represents the state of the articulators at the instantaneous time the phoneme is produced.

Coarticulation is the articulatory (and acoustic) effect which results when one speech segment dynamically influences the production of an adjacent segment. Phonemic coarticulation, common in natural speech, renders the articulatory and acoustic state of a phoneme a variable function of the phoneme which precedes and/or follows. Because the parameter specification for a coarticulated phoneme is no longer unique or "invariant" with respect to its context, coarticulated speech is incompatible with a segmental or static model of production.

From an acoustic standpoint, much of the variability in speech signals can be attributed to coarticulation. It is prevalent enough in natural speech that the task of segmenting a compound utterance into a series of discrete phonemes and boundaries is often elusive and is usually subject to inconsistencies (Cole et al. 1980).

Relative to the analysis of speech sounds in general, therefore, an analysis of coarticulation effects is especially sensitive to the method of segmentation employed by the model. Further, coarticulation effects themselves are generated as a result of the temporal relationships between various articulatory events. For these reasons, the effects of coarticulation are best analyzed, not through a model consisting of segmented phonemes, but through a dynamic or continually time-varying model of the production mechanism. The proposed wavelet model has such time-variability.

1.4 The Speech Effect Model for Vocalic Coarticulation

Consider the speech model illustrated previously in Figure 1.1. Suppose that the control utterance is an isolated vowel articulation, /V/. Let the effected utterance be the same vowel imbedded within a /C−C/ context. A system analysis of the channel which is associated with this operation describes the effect of the context on the vowel. In other words, the channel identifies (in terms of wavelets) the process of CVC coarticulation. The overall system thus models the dynamic transition of the vocal tract from its /V/ articulation to its C/V/C articulation. As expressed in these terms, consonant-vowel-consonant coarticulation exhibited on vowels is illustrated in Figure 1.2:

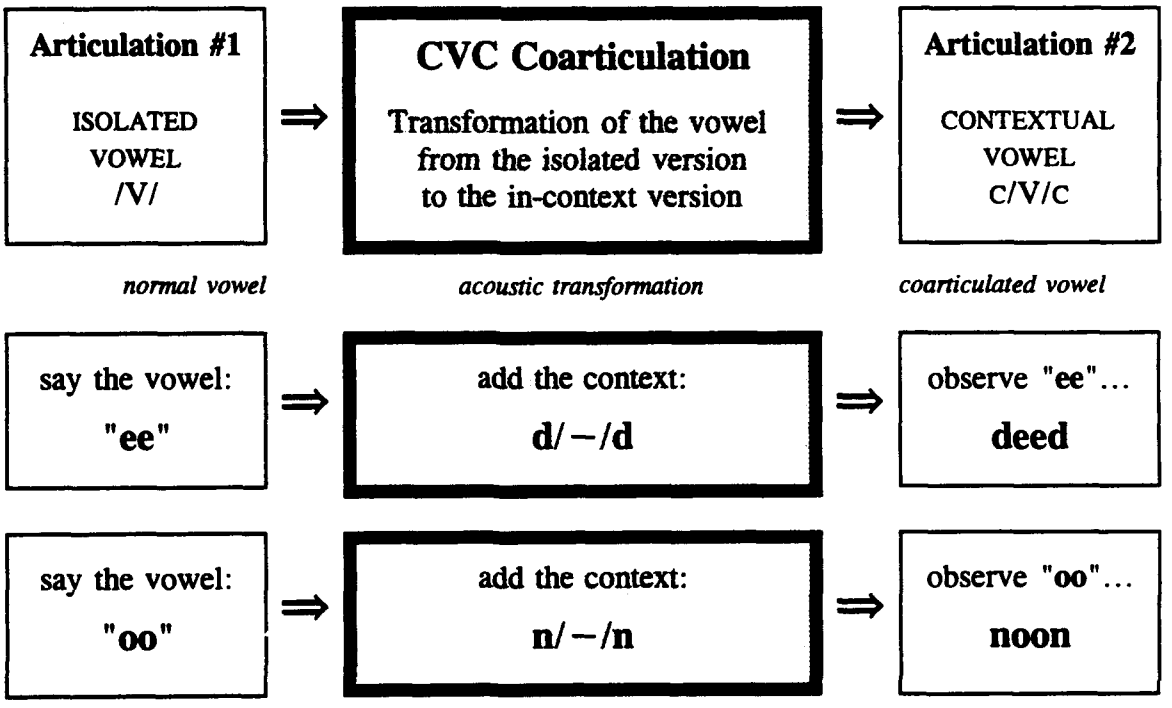


Figure 1.2 **Consonant-Vowel-Consonant Coarticulation in Speech**

1.5 The Objectives of this Research

The goal of this study is to provide a concise acoustic description of consonant-vowel-consonant coarticulation. That is to say, we desire an acoustic description of coarticulation which is sensitive to the phonetic content of the utterance, yet insensitive to the variability due to other sources. A description of coarticulation which is applicable in a variety of phonemic contexts will contribute to our understanding of continuous speech for the following purposes:

1) The clinical description of normal and abnormal productions
2) The synthetic generation of natural sounding speech
3) The computer recognition of natural, unconstrained utterances.

The relevant contribution of this study is two-fold:

1) Propose a theoretical model based on recent developments in wavelets which is capable of describing CVC coarticulation effects on vowels.
2) Evaluate the model by applying it to samples of real utterances.

A practical experiment is conducted whereby samples of real /V/ and C/V/C utterances are processed into working variables for the system model. The effect of CVC coarticulation is then analyzed by measuring the model parameters for a variety of consonant/vowel categories. The model is then evaluated on the basis of how effectively its description of coarticulation reflects phonemic variations from one consonant/vowel category to another.

The purpose of the experiment is to determine whether there exists an invariant phonetic basis for this model's description. In other words, does the *acoustic* description

provided by the model correlate with the *phonetic* parameters of the utterance, such as vowel height and position, or consonant place and manner of articulation? Does the proposed model effectively lower the dimensionality of the CVC coarticulation problem? The experiment is designed to provide evidence for answering these questions.

1.6 Thesis Overview

The thesis is divided into a number of major chapters, each covering a different phase of the research. The Background provides a literature review on the subject of CVC coarticulation. The Theory chapter outlines the basic aspects of wavelet analysis and especially wavelet system theory. The results of that chapter are expressed specifically in terms of variables relevant to speech production. The Model chapter defines, in abstract terms, the speech effect model and the coarticulation model. The Solution is an analysis which derives the parameters of the coarticulation model in terms of practical, measurable quantities. The Experiment chapter describes the methodology for the experimental study. The body of calculations derived from the experiment are presented in the Results, and some aspects pertaining to the verification of these results appear in the Validation. The Conclusion chapter follows. Finally, the Appendix chapters, A and B, contain material concerning some applications of the model and practical issues pertinent to its implementation on real speech.

Chapter 2

BACKGROUND

As far back as the late 1930's, using a harmonic analysis of the speech wave, John Black (1939) concluded that differences in the spectral composition of vowels may be attributed to the effect of the consonants which precede and follow. In the mid 1950's, Carol Schatz (1954) showed that the perception of initial voiceless stops are influenced by their context, specifically, by the vowel which immediately follows. She found this influence, demonstrated using human speech, to be consistent with results previously demonstrated using synthetic speech.

Since the 1950's, much evidence has been collected in support of the locus theory of stop-consonant perception (Cooper et al. 1952; Delattre et al. 1955). Delattre, Liberman, Cooper, and their colleagues contended early in this research that the transition interval from stop-consonant to vowel is characterized by a continuous "movement" of the second formant frequency (*F2*). Specifically, *F2* moves from the value of the locus frequency for the stop to the steady-state *F2* value of the vowel. Evidence of such a pattern resulted from perception studies using synthetic speech, but nevertheless indicates the presence of a contextual influence on a vowel by the adjacent consonant.

With respect to secondary vowel characteristics such as duration, fundamental frequency (*F0*), and intensity, House and Fairbanks (1953) showed the consonant environment to be influential in a systematic way. In particular, they found that the

manner and place of articulation of the consonant within a consonant-vowel-consonant portion of an utterance were factors which significantly affected the duration, *F0*, and intensity of the intermediate vowel.

Stevens and House (1963) specifically examined the coarticulatory effect of adjacent consonants on vowel articulations. They used in their speech sample a two-syllable utterance: /h∩//CVC/, where the initial syllable containing the schwa vowel is unstressed, and the following stressed syllable contains consonant-vowel-consonant. The initial and final consonants in /CVC/ were always the same phoneme. Also included were samples of isolated vowel /V/ articulations. Measurements of the first and second formant frequencies, taken as a time-*average* over the course of the vowel, were made for each vowel utterance. Differences or shifts in these formant frequencies (relative to the isolated /V/ case) were recorded as a function of the context. They found that the effect of the context on vowel formants depended on the consonantal place and manner of articulation. Specifically, changes in the place of articulation corresponded to systematic shifts in *F2*, and changes in the manner of articulation also corresponded to systematic shifts in *F2*. In addition, the magnitude of these effects varied significantly as a function of the vowel.

In his spectrographic studies on vowel-consonant-vowel /VCV/ utterances, Öhman (1966; 1967) demonstrated a number of mutual coarticulatory effects which occur within the vowel-consonant pair. He examined in particular the formant transition interval between vowel and consonant, namely, the transitions between /VC/ and between /CV/. Both stop-consonant and fricative /C/'s were used. He found the time-dynamic shape of the formant transition in /VC/ to be variable and dependent on the final /V/ of the /VCV/

utterance. Likewise, the shape of the formant transition in /CV/ was dependent on the initial /V/ of the /VCV/. The extent of variation attributable to this coarticulation was especially noted in the case where the /C/ was a stop-consonant. In such cases, little or no correlation was found between the terminal frequency of *F2* in the (/VC/ or /CV/) transition and the identity of the stop-consonant /C/. This evidence contradicted some existing theories of invariance which established a strong relationship between the terminal *F2* frequency for the transition and the place of articulation for the stop (locus theory). In the case of the fricative /C/, the observed coarticulation in formant transitions failed to exhibit any such shifts in the terminal *F2* frequency. Thus, an overall contrast between the /VCV/ coarticulation for stops and that for fricatives was observed. The study also indicated that even the "stationary portion" of the vowel was influenced by the identity of the adjacent consonant. Presumably, "stationary portion" refers to the medial portion of the vowel which is distinct from its transition to (or from) the consonant. This influence was observable on vowels occurring in both the initial and final positions of /VCV/. Among the principal conclusions of this study was that the perception of the intervocalic stop must be subject to the entire /VCV/ utterance, rather than to a single invariant cue occurring within one segment. A further interpretation of these results by Öhman was the refutation of an articulation model for /VCV/ which is based on a linear sequence of independent gestures. He maintained that vowel and consonant gestures are "independent" at the level of neural instructions, but not at the level of mechanical articulation.

Stevens et al. (1966) revisited the CVC coarticulation analysis. They performed analyses similar to those reported earlier, namely, measurements of the first and second

formant frequencies (*F1* and *F2*) for the vowel occurring between two identical consonants. The difference in this study, however, was the *dynamic* treatment of the vocalic portion of the /CVC/. Numerous measurements of *F1* and *F2* were made throughout the vowel (each separated by intervals of about 8 ms) using a "spectrum matching" technique. The spectrum matching consisted of an iterative comparison of the sampled spectra with synthetic spectra calculated from acoustic resonator theory. The method results in a pole-zero model of the vocal tract transfer function, expressed as a function of time. The analysis led to the formulation of a coarticulation model featuring a parabolic *F2* trajectory (in time). The ends of the parabola approximate the "loci" values of the consonant (initial and final). The medial value of *F2* achieved either a maximum or minimum (depending on the shape of the trajectory) and approximated the standard *F2* value for that vowel (as in the isolated /V/ case). The analysis for *F1* fit this same pattern, except here the trajectory was consistently concave downward and the parabolic curvature was much less than that of the *F2* trajectory. Specific parameters of the *F2* trajectory, including initial and final frequencies, minimum/maximum frequency, duration, and coefficient of curvature, were used to characterize the influence of the consonant on the vowel. These observed patterns in coarticulation were found to be functions of the vowel features tense/lax and diffuse/non-diffuse and of the consonant feature place of articulation.

Lindblom and Studdert-Kennedy (1967) used synthetic CVC syllables to demonstrate the influence of adjacent consonants on the perception of the vowel. A series of vowel sounds generated from a set of continuously varying formant patterns were inserted between identical initial and final consonants. The vowel sounds ranged

from /u/ to /I/, and two different consonantal frames, /w-w/ and /j-j/, were used. Results indicated that listener categorization of the vowel was influenced significantly by the consonantal environment and by the duration of the vowel. The researchers concluded further that the formant *transition* patterns to and from the consonantal segment (in conjunction with the medial "target" formant values) influenced the perceived identity of the vowel.

Similar conclusions as to which factors contribute to vowel perception were reached by Strange et al. (1976) using human speech. They showed that vowels produced in a /p-p/ environment were identified by listeners with much greater accuracy than their counterparts spoken in isolation. In a second experiment, the consonantal context was varied unpredictably, and the vowels appearing in these environments were still identified with greater accuracy than those spoken in isolation. These results led to the overall conclusion that listeners utilize dynamic acoustic information over the entire duration of the vowel; no single time slice or time-average is sufficient to specify the acoustic and perceptual properties of the vowel occurring in /CVC/.

Chapter 3

THEORY

3.1 The Classical Model

The standard acoustic model of speech production is the source-filter model (Fant 1960, section 1.11):

$$Z(\omega) = H(\omega) \cdot G(\omega)$$

where, for the restricted class of vocalic speech sounds, $G(\omega)$ is the glottal source spectrum, $H(\omega)$ is the transfer function of the vocal tract (including the nasal cavity), and $Z(\omega)$ is the speech (output) spectrum. The variables Z, H, and G are functions of frequency (ω). The equivalent time-domain expression is:

$$z(t) = h(t) \odot g(t)$$

where $g(t)$ is the glottal source (excitation) signal, $h(t)$ is the vocal tract impulse response, $z(t)$ is the acoustic pressure at the point of a microphone transducer, and the operation $\odot$ denotes convolution. Variables z, h, and g are functions of time (t).

As a model describing speech production, an evaluation of this model's parameters determines which one of a variety of possible vocalic speech sounds is generated. In particular, an evaluation of the vocal tract parameter $h(t)$ specifies the articulation of a specific vocalic phoneme.

The above relation is described in general signal terms as a system operator (convolution) acting on the input [the glottal source $g(t)$] under some parameterization of the channel [the vocal tract impulse response $h(t)$]. The output [the pressure $z(t)$ at the transducer] is a result of the operation on the input. Acoustically, the vocal tract channel acts as a linear filter (Flanagan 1972, chapter 3). In other words, the output spectrum $Z(\omega)$ is a filtered version of the input spectrum $G(\omega)$, as specified by the transfer function $H(\omega)$.

The classical model is a steady-state description of vocalic voice production. The time series g and z are assumed to be stationary. The impulse response h is the response over all time to an impulse which occurs at a single instant in time. The convolution operation is taken over the entire life of the input signal (from time equals minus-to-plus infinity). The power spectra G, H, and Z, therefore, describe resonances which are *in*variant with respect to time, hence the characterization of the system as Linear Time-Invariant (LTI). Figure 3.1 illustrates the Linear Time-Invariant system:

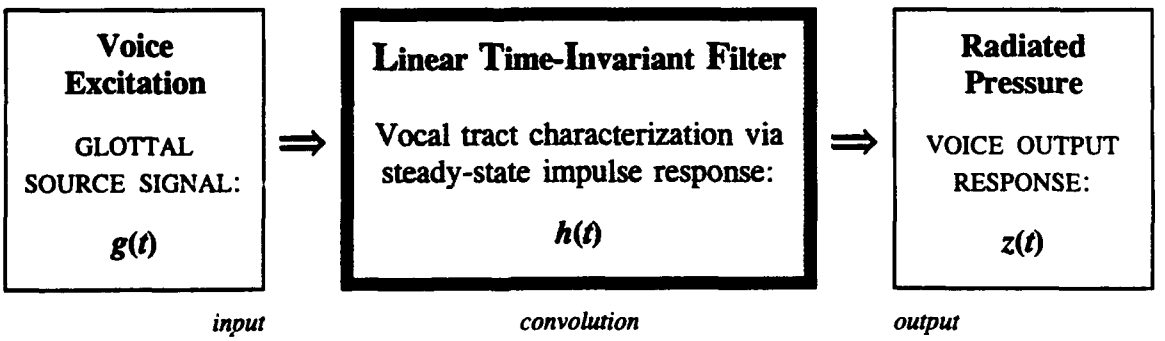


Figure 3.1 **The LTI Source-Filter Model for Vocalic Speech Production**

3.2 The Wavelet Transform as a Signal Analysis Tool

Wavelet transforms are typically used as a method of analysis for time-varying signals. The transform generates a second-order representation of a signal. It expresses the signal as a weighted sum of wavelet components, distributed as a function of time and scale. In many cases, the scale parameter is akin to frequency (Daubechies 1990, section I). A wavelet description of a signal is like the classical spectrogram, in that it provides information about the content of the signal with respect to its constituent scale-values (frequencies) and their presence or absence as a function of time.

The wavelet transform has the following advantages for speech:

1) A wavelet representation is non-parametric. By this it is meant that no specific model constraint or a priori form for the process is employed. For example, in a linear predictive analysis, the form of the model is based on the location and magnitude of a finite number of spectral peaks. Wavelet analysis, on the other hand, assesses the relative magnitudes of all signal components, regardless of their proximity to the "peaks".

2) The time-frequency resolution of wavelet analyzers varies linearly along the frequency spectrum. Specifically, at high scales values, the resolution bandwidth is broad and the time-resolution superior. At low scale values, the resolution bandwidth is narrow and the time-resolution long. This distribution of the time-frequency resolution reflects the distribution of energy (in time and frequency) for most speech sounds. For example, high frequency stop-burst noise typically occupies a wide

bandwidth over an extremely short time interval. On the other hand, low frequency vowel formants exhibit narrow-band resonances over relatively longer time durations.

3) No assumptions are made about the short-term stationarity of the signal.

The wavelet transform is defined as (Grossmann et al. 1989):

[3.1] $$\mathbf{W}_{f(t)} x(t) \quad (a,b) \quad = \quad \frac{1}{\sqrt{|a|}} \int x(t)\, f^{*}\left(\frac{t-b}{a}\right) dt$$

where:

$x(t)$	is the function under analysis/transformation.
$f(t)$	is the analyzing "mother" wavelet function.
a,b	are the time-scale and time-shift parameters, respectively.
$\int dt$	represents an integral from minus to plus infinity.
f^{*}	denotes the complex conjugate operation on the function f.
$\mathbf{W}_{f} x \ (a,b)$	denotes the wavelet transform of $x(t)$ with respect to the wavelet $f(t)$, using a,b as the time-scale and time-shift parameters.

The aspect of equation [3.1] which is central to the structure of the wavelet transform is the "affine mapping" of the mother wavelet (Young 1993, chapter 1). The affine mapping operation refers to the shifting (b) and scaling ($^{1}/_{a}$) of the wavelet function (f), with respect to time (t):

[3.2] $$f_{a,b}(t) = f\left(\frac{t-b}{a}\right)$$

$f_{a,b}$ represents the family of wavelets which appear as shifted and scaled versions of the mother wavelet, $f(t)$. Figure 3.2 shows some examples from one such family. The mother function is the Morlet wavelet, $f_M(t)$:

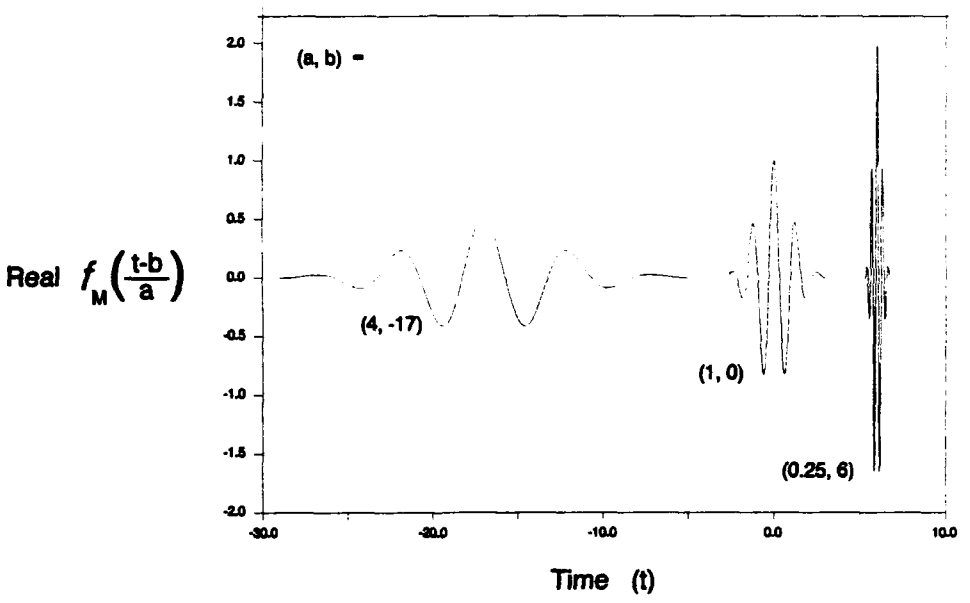


Figure 3.2 **Shifted and Scaled Versions of the Morlet Wavelet** $f_M(t) = e^{j5t} \cdot e^{-(t \cdot t)/2}$

The function in the figure which corresponds to $(a,b)=(1,0)$ is equivalent to the mother wavelet; it has unity scale and zero time-shift. The wavelet $(a,b)=(4,-17)$ depicts a dilated version of the mother which is shifted earlier in time. In contrast, the wavelet $(a,b)=(0.25,6)$ is a compressed version of the mother which is delayed in time.

One interpretation of the wavelet transform $\mathbf{W}_f x\ (a,b)$ is:

an analysis of the function x(t) in terms of the wavelet f(t).

In other words, $\mathbf{W}_f x\ (a,b)$ represents $x(t)$ as a scaled and shifted version of $f(t)$. Alternatively, the transform $\mathbf{W}_f x\ (a,b)$ may be viewed as:

the correlation of x(t) with f(t)

-on the basis of two parameters: the scale parameter (a) and the time-shift parameter (b).

The mother wavelet function $f(t)$ need not be a windowed sinusoid, as is the case for the Morlet. Rather, a variety of time-functions may be used as the mother wavelet. Different mother wavelets yield different transform coefficients $[\mathbf{W}_f x\ (a,b)]$ for a wavelet transform of the same signal. More specifically, each mother wavelet *derives its own basis* for depicting the content of a given signal.

3.3 The Wavelet Transform as a System Analysis Tool

In the doctoral thesis of Randy K. Young entitled "Wideband Space-Time Processing and Wavelet Theory" (1991), the method of wavelet system characterization is introduced. Wavelet system characterization provides a quantitative description for the behavior of a transmission channel, expressed in terms of the channel's input and output signals. This method assumes that the system's input and output signals are time-*varying*, and that the system transmission channel is itself time-varying.

The operator used for characterizing a system in this fashion is called the Space-Time Varying [STV] operator (Young 1991, chapter 5, section 5.3.1). The STV

operator uses the system input function to generate the output, but it also incorporates another function which specifically represents the behavior of the *channel*. This channel function is a two-dimensional distribution of wavelet-coefficients.

The **STV** *channel characterization* is a wavelet transform which depicts how the input function may be scaled and time-shifted in order to yield the output function. In particular, the **STV** channel characterization is estimated as the wavelet transform of the system *output* signal, using the system *input* signal as the mother wavelet. In other words, the channel characterization consists of wavelet coefficients of time-scale (a) and time-shift (b). These coefficients serve as a weighting-function which, when applied to the input, reproduce in the output the transformation effect of the channel.

A speech production model which employs the **STV** operator is viable for the following reasons:

1) The **STV** operator models a transmission channel which is linear, just as the classical LTI (linear time-invariant) system models a filter which is linear (Bendat and Piersol 1986, chapter 2).

2) The **STV** operator includes a parameterization of the channel which is time-varying, rather than time-invariant or steady-state. Thus, no assumptions are made about the short-term stationarity of either the signal or the channel. In terms of speech, no time-segmentation is employed by either the model or the process of analysis. A more reliable characterization of the transient events in speech is therefore possible.

3) The **STV** channel parameterization is a wavelet-type description of the channel. Therefore, the same advantages afforded through the use of

wavelets as a method of signal analysis are also afforded for wavelet system analysis.

The **STV** operator for modeling a system appears as follows (Young 1991, chapter 5, equation 5.8):

[3.3] $$\mathbf{STV}_{P(a,b)}[x(t)] = y(t)$$

where:

$x(t)$	is the input to the channel.
$y(t)$	is the output of the channel.
a,b	are time-scale and time-shift parameters.
$P(a,b)$	is a representation of the channel. It describes the channel behavior in terms of wavelet transform-domain coefficients.

The **STV** operates on $x(t)$ under $P(a,b)$, and the result is $y(t)$. By virtue of the time-shift parameter b, the channel representation $P(a,b)$ is a dynamic function of time.

The structure of the **STV** operator, and its capacity for describing a system, is apparent when posed in terms of the identification of parameter $P(a,b)$. Just as the $P(a,b)$ channel characterization is expressed in terms of wavelet-domain coefficients, it can be *estimated* in terms of a wavelet transform. As specified by Young (1991, chapter 5, equation 5.15), this estimate appears thus:

[3.4] $$\hat{P}(a,b) = \mathbf{W}_{x(t)} y(t) \quad (a,b)$$

where $\hat{P}$ denotes an estimate of the function P. Equation [3.4] depicts the channel characterization $[\hat{P}(a,b)]$ as the wavelet transform of the system output $[y(t)]$, using the system input $[x(t)]$ as the analyzing wavelet function.

In accordance with the wavelet transform interpretation which appears on page 19, $\hat{P}(a,b)$ is the correlation of the output with the input. (For a more detailed explanation of this interpretation, see page 38). The $\hat{P}(a,b)$ channel estimate appearing in [3.4] can also be viewed as the analysis of the output signal $[y(t)]$ in terms of the input signal $[x(t)]$.

From the point of view of equation [3.3], the **STV** operator performs a scaling and shifting of the input. The result $y(t)$ is a weighted sum of scaled and shifted versions of $x(t)$, as dictated by the wavelet-coefficient distribution $P(a,b)$. In particular, $x(^{t-b}/_{a})$ is weighted by $P(a,b)$ and summed over many values of scale (a) and shift (b). The following double integral shows this weighted sum (Young 1993, chapter 5, equation 5.11):

$$\textbf{[3.5]} \qquad y(t) = \underset{P(a,b)}{\textbf{STV}}\,[x(t)] = \int \frac{1}{a^2} \int P(a,b)\, \frac{1}{\sqrt{|a|}}\, x\left(\frac{t-b}{a}\right) db\, da$$

Equation [3.5] thus expresses explicitly the **STV** operation of equation [3.3].

As previously stated, the **STV** channel is like the LTI filter in that the operation is linear. Unlike the **STV** channel, however, the LTI filter is time-invariant. *Time-invariant* linearity ensures that the filtering behavior of the operation consists of some magnitude and phase adjustment for each input spectral frequency. Any spectral components which do *not* appear in the LTI input, however, cannot be "generated" by

the filter. In other words, input frequencies may be amplified by an LTI filter, but no "new" frequencies can appear in the output which were not already present in the input.

In contrast, the **STV** channel *does* map "new frequencies" to the output. The scaling parameter a specifically designates a time-warping of the input. In particular, $a < 1$ effects time-dilation, and $a > 1$ effects time-compression. The **STV** operator models a channel which is thus capable of generating (for any value of $a \neq 1$) frequency transitions (frequency movements) which vary as a function of time.

Equation [3.4] shows how the **STV** operator and the wavelet transform may be used to estimate the characteristics of a system's channel. The use of the **STV** operator in a model for vocalic speech, therefore, provides a description specifically for the behavior of the vocal tract *channel*. This is shown in the following section.

3.4 STV Parameters for the Vocal Tract

Let a vocalic speech utterance be modeled according to the following definitions:

$n(t)$ ≡ broadband $1/f$ noise source.

$r(t)$ ≡ the vocal tract noise response. It is the output pressure when the vocal tract is excited by broadband $1/f$ noise.

$g(t)$ ≡ glottal source time function.

$z(t)$ ≡ the voice output response measured at a microphone transducer when the vocal tract is excited by $g(t)$.

$P(a,b)$ is the **STV** wavelet-coefficient representation of the vocal tract. a,b are the time-scale and time-shift parameters.

The following subscripts apply:

r1, *g1*, *z1*, *P1* : are objects of one vocalic utterance; #1

r2, *g2*, *z2*, *P2* : are objects of a *different* vocalic utterance; #2

Utterances #1 and #2 are distinct. None of their respective quantities are assumed to be equal.

The broadband $^1/_f$ noise source is a method for exciting the vocal tract channel in a manner which yields (via the vocal tract noise response) a complete description of the time-frequency behavior of the channel. In this respect, $^1/_f$ noise does for the **STV** wavelet model what the unit impulse excitation does for the LTI filter model. Whereas the unit impulse excites the LTI channel with equal power at all frequencies, $^1/_f$ noise excites the **STV** channel with equal energy at all time and frequency locations. More specifically, the *spectral density* of $n(t)$ is a function which decays as $^1/_f$ (hence the name). The *wavelet* transform of $n(t)$ generates a time-frequency distribution of wavelet coefficients which is constant. In a sense, $^1/_f$ noise is white noise for wavelet analysis (Fowler 1991; Wornell 1990).

The quantities defined above can be used in an **STV** system to describe vocalic speech production in terms of wavelet parameters. Consider first a noise excitation of the vocal tract, resulting in a noise response output. According to the definition of the **STV** operator, equation [3.3], the system appears:

$$\mathbf{STV}_{P1(a,b)}[n(t)] = r1(t) \qquad \mathbf{STV}_{P2(a,b)}[n(t)] = r2(t)$$

$n(t)$ input under $P(a,b)$ yields $r(t)$ output.

In $P1(a,b)$, the vocal tract assumes the shape of one articulation. In $P2(a,b)$, the vocal tract assumes a different shape. In either case, the same noise $[n(t)]$ is input. $r1(t)$, therefore, is the noise response generated from the first vocal tract articulation. Likewise, $r2(t)$ is the noise response generated from the second vocal tract articulation. Figure 3.3 illustrates the Space-Time varying channel associated with the noise-excited vocal tract:

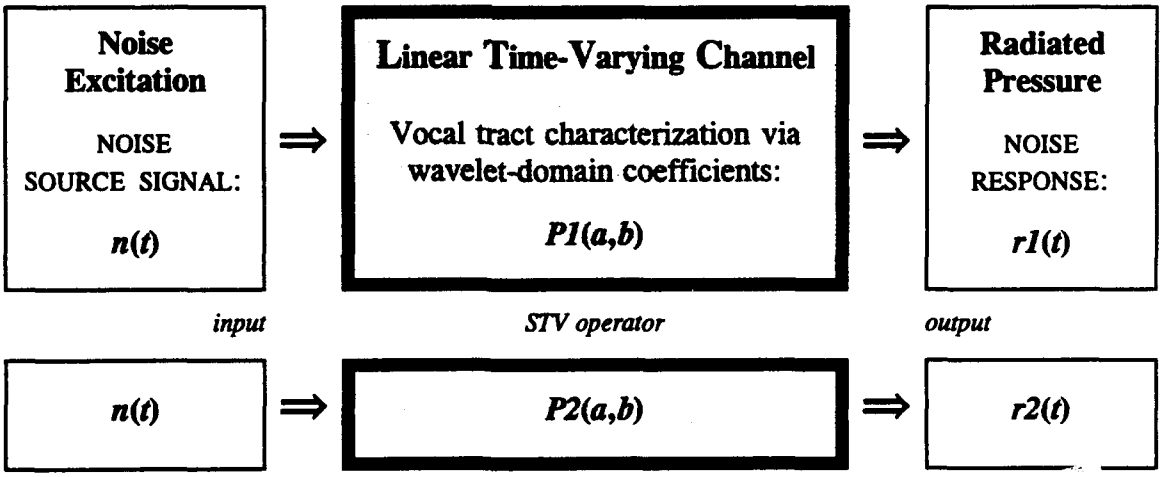


Figure 3.3 **The STV Model for the Noise-Excited Vocal Tract**

Consider next a *glottal* excitation of the vocal tract, resulting in a real utterance output. The vocal-tract articulations (*P1* and *P2*), however, are maintained. The **STV** system appears:

$$\mathbf{STV}_{P1(a,b)}[g1(t)] = z1(t) \qquad \mathbf{STV}_{P2(a,b)}[g2(t)] = z2(t)$$

$g(t)$ input under $P(a,b)$ yields $z(t)$ output.

In $P1(a,b)$, the vocal tract assumes the same shape as in the noise case. Likewise, $P2(a,b)$ is the same as in the previous case. However, $z1(t)$ represents a real utterance, the result of a glottal excitation $g1(t)$ channeled through the vocal tract $P1(a,b)$. Similarly, $z2(t)$ is another utterance, the result of a different glottal excitation $g2(t)$ channeled through a different articulation $P2(a,b)$.

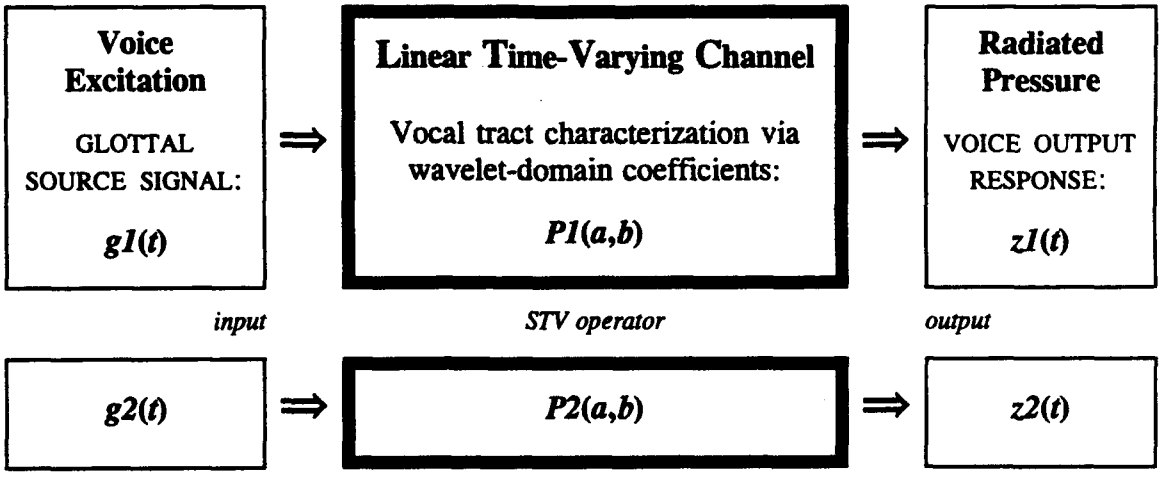


Figure 3.4 **The STV Model for a Real Vocalic Utterance**

Notice how the **STV** system model appearing in Figure 3.4 contrasts with the LTI system (Figure 3.1).

As stated previously, the **STV** channel characterization is estimated in terms of a wavelet transform. According to equations [3.3] and [3.4], this estimate for $P(a,b)$ appears as the wavelet transform of the **STV** *output* with respect to the *input*. As they appear in the two systems above, therefore, estimates for $P1$ and $P2$ can be formulated:

$$\hat{P1}\ (a,b) = \mathbf{W}_{n}\ r1\ \ (a,b) \qquad \hat{P2}\ (a,b) = \mathbf{W}_{n}\ r2\ \ (a,b)$$

[3.6]

$$\hat{P1}\ (a,b) = \mathbf{W}_{g1}\ z1\ \ (a,b) \qquad \hat{P2}\ (a,b) = \mathbf{W}_{g2}\ z2\ \ (a,b)$$

Equation [3.6] gives a pair of estimates for $P1$, which is the wavelet characterization for one articulation of the vocal-tract. The first estimate for $P1$ is expressed in terms of a noise source input (n) and the resulting noise response ($r1$). The second estimate for $P1$ is expressed in terms of a real glottal source input ($g1$) and the resulting measured output ($z1$). A pair of estimates for $P2$ is likewise stated. $P2$ is the wavelet characterization for a different articulation of the vocal-tract.

3.5 The Mother Mapper Formulation

In his dissertation, Young (1991, chapter 4) introduces another wavelet relation called the "mother mapper". In general, the mother mapper provides a mapping from

one wavelet transform to another wavelet transform. In each of these wavelet transforms, the function under transformation is the same. However, the mapping occurs between a wavelet transform using one analyzing wavelet and a wavelet transform using a different analyzing wavelet. (A transform under one "mother" wavelet is mapped into a transform under another "mother" wavelet, hence the name "mother mapper".)

$$\textbf{Mother Mapper:} \qquad \mathbf{W}_{f1}x \quad \Rightarrow \quad \mathbf{W}_{f2}x$$

Specifically, $\mathbf{W}_{f2}x$ is expressed as a function of two other wavelet transforms:

[3.7] $\mathbf{W}_{f2}x$ = an integral function of $\mathbf{W}_{f1}x$ and $\mathbf{W}_{f1}f2$

The motivation for reformulating the wavelet transform according to [3.7] is seen by considering the affine operation of the mother wavelet *(f)* which appears in equation [3.2]. Typically, when a standard analyzing function is employed as the mother wavelet, f is expressed analytically. It is thus well defined at all possible values of $(t-b/a)$. Suppose, however, that the mother wavelet is instead a measured random signal (such as that obtained from a sample of human speech). Let this "measured" wavelet function be denoted $y(t)$. Re-expressing the wavelet transform definition in [3.1] yields:

$$\mathbf{W}_{y(t)}x(t) \quad (a,b) \quad = \quad \frac{1}{\sqrt{|a|}} \int x(t)\; y^{*}\left(\frac{t-b}{a}\right) dt$$

The wavelet $y(t)$ is a random time series.

Implementation of this wavelet transform requires that $y(t)$ be "measured" or sampled at each time value given in $y(^{t-b}/_{a})$. Because the scale-value (a) assumes integer as well as non-integer values, no *regular* sampling interval *(T)* exists, such that $(^{t-b}/_{a})$ always equals an integer multiple of T. In other words, $y(^{t-b}/_{a})$ requires knowledge of $y(t)$ at numerous intermediate time-values outside of those regular intervals captured by a uniform sampling rate. The standard method of digital signal sampling is thus inadequate for obtaining $y(^{t-b}/_{a})$, unless an extremely high order of over-sampling is employed, or unless approximations to $y(^{t-b}/_{a})$ are calculated. The order of over-sampling and/or approximation to $y(^{t-b}/_{a})$ required for the purposes of speech render the straightforward implementation of this wavelet transform impractical.

One of the primary advantages of the mother mapper formulation, therefore, is a resulting method for performing a wavelet transform on a measured random signal [$x(t)$] with respect to a measured random wavelet [$y(t)$], without the need to scale *either* function. Using the mother mapper, $\mathbf{W}_y x$ can be derived from the wavelet transforms of each $x(t)$ and $y(t)$ *individually*.

The following relation shows this mapping for $\mathbf{W}_y x$. Using equation [3.7], a random time series $y(t)$ is substituted in place of the wavelet function *f2(t)*. Another wavelet [*f(t)*] is used as the common analyzing wavelet:

[3.8]

$$\textbf{Mother Mapper:} \quad \mathbf{W}_f x \Rightarrow \mathbf{W}_y x$$

$$\mathbf{W}_y x = \text{an integral function of } \mathbf{W}_f x \text{ and } \mathbf{W}_f y$$

The wavelet $y(t)$ is a random time series. The wavelet *f(t)* is an analytic function.

Notice that each of the "functional" wavelet transforms (on the right-side of this equation) is made with respect to the *same* wavelet $f(t)$. In this formulation, $f(t)$ is to be chosen as a standard wavelet function, expressed analytically and therefore "known" at all time-values. The scaling of $f(^{t-b}/a)$, which is a necessary operation for a wavelet transform construction, thus remains a simple matter (see equation [3.2]).

The above relation is expressed explicitly in the following equation. According to the results of Young (1991, chapter 4, equation 4.28), the mother mapper integral appears:

$$[3.9] \qquad \mathbf{W}_{y}\, x\,(s,\tau) \;=\; \frac{1}{C_f}\int \frac{1}{a^2}\int \mathbf{W}_{f}\, x\,(a,b)\cdot \mathbf{W}^{*}_{f}\, y\left(\frac{a}{s}\,,\,\frac{b-\tau}{s}\right) db\; da$$

where:

$x(t)$, $y(t)$	are (both) random time series.
s,τ	are the wavelet transform time-scale and time-shift parameters, respectively.
C_f	is a normalizing constant for $f(t)$.
$f(t)$	is an analytic wavelet function.
$\mathbf{W}^{*}$	denotes the complex conjugate operation on the wavelet transform $\mathbf{W}$.
$\mathbf{W}\,(^{a}/_{s}\,,\,^{b-\tau}/_{s})$	denotes the scaling $(^{1}/_{s})$ and shifting $(b-\tau)$ of the parameters a,b in the wavelet transform $\mathbf{W}$.

Notice that the functional wavelet transforms $\mathbf{W}_f x$ and $\mathbf{W}_f y$ are formulated in terms of the scale and shift parameters a,b. However, one of these transforms $(\mathbf{W}_f y)$ is itself

scaled and shifted with respect to the other $(\mathbf{W}_f x)$. These scales and shifts in $\mathbf{W}_f y$ are effected through the scale factor s and the time-shift τ. Integration occurs over the pair a,b, so that the *resulting* wavelet transform $(\mathbf{W}_y x)$ is a function of s,τ. (This construction is reminiscent of the standard convolution integral, which effects similar shifts along a single parameter.)

Chapter 4

MODEL

The previous analysis is a re-expression of results from Young[7] stated in terms of the parameters involved in voice production. The analysis which follows is an extension on that framework. The result constitutes the proposal of a wavelet model for vocalic speech coarticulation.

4.1 STV Channel as a Speech-Effect Transfer

To this point, the system embodying the **STV** operator is considered as a physically realizable process. This means that the system input, channel, and output are assumed to exist within physical dimensions. They are linked in space by a series: the input signal excites the channel which, in turn, outputs its response. In the case of vocalic speech, for example, a glottal source input propagates through one end of the vocal tract channel. The response, radiated from the opposite end, is the corresponding vowel sound.

Consider next, however, the **STV** system as a transition through *speech* states. In other words, let an **STV** operator perform the transformation or mapping from one speech condition to another. The input to the system might be an isolated control utterance. The system output would be the same utterance, but effected. The effect

could be a particular voice quality, the influence of a particular phonetic context, or the influence of the speaker.

Such an **STV** system can be expressed in the same terms used previously, *if* each of the input/output states are embodied in an utterance and parameterized according to some time function. Let the time function which characterizes an utterance, therefore, be called the waveform for the utterance, and let it be denoted $w(t)$. Because the speech effect which distinguishes the input and output states may assume a variety of forms, $w(t)$ might correspond to a complete compound utterance or one particular part of an utterance.

The **STV** speech operator therefore describes a transitional mapping from one speech state to another, whereby, the initial (control) state is defined by the waveform of one utterance [$w1(t)$]. The effected state is manifested (relative to the first) by the waveform of a *different* utterance [$w2(t)$].

The **STV** speech-effect system is expressed below in analytical form. Beginning with the general structure of the **STV** (stated in equation [3.3]):

$$\mathbf{STV}_{P(a,b)}[x(t)] = y(t)$$

The **STV** operates on $x(t)$ under $P(a,b)$, and the result is $y(t)$.

The model for the speech-state transition thus becomes:

[4.1]

$$\mathbf{STV}_{P_{SE}(a,b)}[w1(t)] = w2(t)$$

where:

$w1(t)$ is the waveform corresponding to the initial or control speech state.

$w2(t)$ is the waveform corresponding to the final or effected speech state.

a,b are the time-scale and time-shift wavelet parameters.

$P_{SE}(a,b)$ is a characterization of the speech effect. It describes a speech transformation, i.e., a mapping from the control state to the effected state.

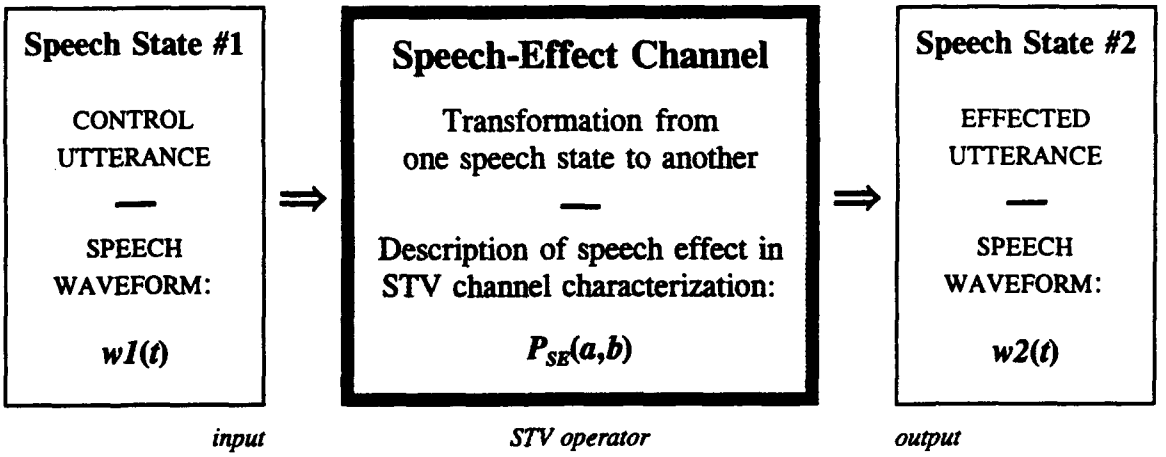


Figure 4.1 **The Speech Waveform Channel**

As illustrated in Figure 4.1, equation [4.1] formulates the STV operator in terms of a speech-state transformation. The "operation" functions abstractly, by means of generating a particular speech effect from a given input waveform. As for any STV

channel characterization, $P_{SE}(a,b)$ is defined in terms of wavelet transform-domain coefficients.

Notice the inherent structure of this speech-effect model, namely, the formulation of a *comparison* between two utterances. An utterance is not characterized absolutely by the $P_{SE}(a,b)$; rather, one utterance is characterized *relative* to a another. Such a characterization provides a direct description of how they differ. Whereas, a more traditional method of characterizing a speech effect might require two stages of analysis (one for the effected utterance and another for the control utterance); the present model documents the difference or "transfer" (from control to effected) within a single, unified description.

4.2 Example Applications of the P_{SE} Speech-Effect Model

If it can be suitably defined in **STV** system terms, some speech effect may be modeled by an $P_{SE}(a,b)$ channel. The following evaluations identify some specific examples of how a speech effect might be modeled by the speech waveform channel:

1) Let $w1(t)$ be the waveform of an isolated "oral" vowel. Let $w2(t)$ be the waveform of the same vowel nasalized. $P_{SE}(a,b)$ may then describe, for that vowel, the effect of *nasality*.

2) Another voice-quality effect might be characterized by setting $w1(t)$ as before and letting $w2(t)$ assume the waveform of a vowel spoken with

twang (Steinhauer et al. 1992). The associated $P_{SE}(a,b)$ would be a characterization of the twang vocal quality in that utterance.

3) Let $w1(t)$ be the waveform of an utterance produced by a speaker without apparent dialect markers. Let $w2(t)$ be the waveform of the same utterance produced by a different speaker with an apparent *accent* or *dialect*. The differences between $w1(t)$ and $w2(t)$ are reflected in $P_{SE}(a,b)$, which becomes a representation of the accent or dialect in $w2(t)$.

4) Another speaker-related effect might be parameterized in $P_{SE}(a,b)$ by letting $w1(t)$ be the waveform of an utterance produced by a male and $w2(t)$ be the waveform of the same utterance produced by a female. The distribution $P_{SE}(a,b)$ then functions as a *gender* transformation for that utterance, whereby, the particular speakers serve as the "prototype" speakers for their respective genders.

5) Suppose that $w1(t)$ is the waveform associated with an isolated phoneme. $w2(t)$ is the waveform associated with the same phoneme spoken by the same speaker, but it is produced within a phonetic *context*. As the transitional mapping from $w1(t)$ to $w2(t)$, therefore, $P_{SE}(a,b)$ describes the effect of the context on the phoneme. In other words, $P_{SE}(a,b)$ identifies segmental coarticulation, or how the production of a phonetic segment is influenced by its adjacent segments (relative to its production in isolation).

This group is not intended to be an exhaustive list of implementable speech effects. Rather, these examples are intended to demonstrate how the P_{SE} system might function in a variety speech-effect applications, including effects in voice quality, speaker effects, and coarticulation effects.

4.3 Estimating the P_{SE} for a Given Speech Effect

As shown for previous STV channel descriptions, the speech-effect characterization in $P_{SE}(a,b)$ can be estimated by a wavelet transform. According to equation [3.4], the estimate appears as the wavelet transform of the STV system output with respect to the input. Using equation [4.1], therefore:

$$[\hat{P}_{SE}]\ (a,b) = \mathbf{W}_{w1(t)}\ w2(t)\ \ (a,b) \qquad [4.2]$$

where $[\hat{P}_{SE}]$ denotes an estimate of P_{SE}. Thus, to the extent that the waveforms $w1(t)$ and $w2(t)$ highlight a speech effect in a representative way, $[\hat{P}_{SE}](a,b)$ *estimates* the appropriate STV channel distribution for that effect.

An alternative interpretation of the $[\hat{P}_{SE}](a,b)$ recognizes the wavelet transform $\mathbf{W}_{w1}w2$ as a representation of $w2$ in terms of $w1$. (This interpretation appears on page 19). $[\hat{P}_{SE}](a,b)$, or the wavelet transform of $w2$ with respect to $w1$, describes $w2$ as a *scaled* and *shifted* version of $w1$. By virtue of the speech effect associated with $w2$, therefore, the control $w1$ is scaled and shifted according to the prescription $[\hat{P}_{SE}](a,b)$.

$[\hat{P}_{SE}](a,b)$ may be interpreted additionally as a correlation function. Consider that the wavelet transform definition (equation [3.1]) contains an integral/product structure which is reminiscent of the correlation integral. As such, $\mathbf{W}_{w1}w2$ may be viewed as a correlation between *w2* and *w1*. Using [4.2]:

$$[\hat{P}_{SE}](a,b) = \mathbf{W}_{w1(t)} w2(t) \quad (a,b)$$

$$= \frac{1}{\sqrt{|a|}} \int w2(t)\ w1^{*}\left(\frac{t-b}{a}\right) dt$$

In the above expression, an inner product occurs between *w2* and *w1*. The integral is formed over the variable *t* of which *w2* and *w1* are functions. One of the functions (*w1*) is further parameterized in *a,b*. *a* and *b* thereby form the basis of correlating *w2* with *w1*. In other words, a correlation is formed over *two* parameters, time-scale and time-shift. In short, $[\hat{P}_{SE}](a,b)$ provides a correlative comparison of two different utterances.

The statistics (distribution, mean, and variance) associated with an $[\hat{P}_{SE}](a,b)$ estimate are not known. Many such estimates of the "true" $P_{SE}(a,b)$ could be derived, however, from an ensemble of "instances" of the $w1(t),w2(t)$ waveform-pair. The $[\hat{P}_{SE}](a,b)$ is an *un*biased estimator. It is expected that the generality of a $[\hat{P}_{SE}](a,b)$ *mean* is defined by the scope of these constituent $w1(t),w2(t)$ pairs.

4.4 Employing the P_{SE} for Synthetic Waveform Generation

The proposed speech-effect transfer model is formulated for identification of the STV channel parameter $P_{SE}(a,b)$. $P_{SE}(a,b)$ is the characterization of some speech effect, and its identification is given in terms of the estimate of equation [4.2]. As stated previously, $P_{SE}(a,b)$ can potentially be used as the parameterization for any number of speech effects, including voice quality, speaker differences, and coarticulation. Once, identified, however, $P_{SE}(a,b)$ could be utilized for an inverse function. Given a normal or control version of an utterance as the input, $P_{SE}(a,b)$ *generates* the effected version of that utterance.

This is shown by observing the specific STV operator function of equation [3.5]. Input $w1(t)$ is substituted for $x(t)$, output $w2(t)$ is substituted for $y(t)$, and $P_{SE}(a,b)$ is used as the $P(a,b)$ channel characterization:

$$[4.3] \qquad w2(t) = \underset{P_{SE(a,b)}}{\mathbf{STV}} [w1(t)] = \int \frac{1}{a^2} \int P_{SE(a,b)} \frac{1}{\sqrt{|a|}} w1\left(\frac{t-b}{a}\right) db\, da$$

$w2(t)$ is the effected version of the control utterance $w1(t)$, as prescribed by the effect-characterization in $P_{SE}(a,b)$.

This synthetic $w2(t)$ generation which is shown in equation [4.3] might be utilized in the development of a coded synthetic speech data base. A bank of control phones (which are contextually isolated) is combined with the $[\hat{P}_{SE}]$ estimated for a particular effect (which is the same for each phone). The generated output is a series of natural

sounding synthetic phones, each possessing the acoustic attributes appropriate for that effect or context.

In a speech *recognition* application, the same method could be used in reverse. A sample of real speech (naturally spoken in context) could be inverted through the P_{SE} to yield a basic isolated version of the phone. This "prototype" version is then fed into the normal recognition stages, but at a level of variability which is, consequently, much reduced. The necessary recognition comparisons between phones could then be performed solely on the basis of *phonetic* discrimination, whereby, any differences due to *contextual* effects have been "removed". The specific formulation for this inversion of the P_{SE} channel is given in the appendix (page 165).

These applications of the P_{SE} model rely on a critical assumption. The assumption is that the $P_{SE}(a,b)$ distribution for a well-defined speech effect is, in fact, *constant* for every phone. It may be, instead, that the $P_{SE}(a,b)$ for one effect is a function of the phonetic context (vowel class, place of articulation, etc.) or of the speaker and the variables associated with his or her voice. It is not known whether there is *any* basis for which $P_{SE}(a,b)$ is an invariant representation of a given speech effect. The purpose of the proposed experiment is to identify the existence or non-existence of an invariant basis for the $P_{SE}(a,b)$ associated with one particular speech effect. The speech effect to be examined in this respect is outlined in the following section.

4.5 The P_{SE} Model for the Coarticulation Effect

The previous section outlines some examples of speech effects which might be successfully modeled by an P_{SE} channel. One of these examples addresses the effect of coarticulation exhibited on a segment by virtue of its phonetic context (enumerated fifth on page 36). Consider a special case of this example, whereby, the phonetic class of the segment is specified, and its waveform signal representation is defined.

Let the input to the channel be an isolated vowel, and the output be the same vowel imbedded within a consonant-vowel-consonant (CVC) context. Assume that the initial and final consonants are the same, and that the utterances are produced by the same speaker. The **STV** channel associated with this system describes the effect of the /C−C/ context on the vowel, /V/. The channel models the acoustic effect of CVC coarticulation.

Under these constraints, the input and output utterances may be represented in the signal terms which are most appropriate to aspects of their articulation. Specifically, both of the vowels in /V/ and C/V/C are vocalic utterances. As such, the dynamic shape of the vocal tract becomes the articulatory component which distinguishes them from each other and from other vowels. Therefore, let /V/ and C/V/C be represented in signal terms by their associated vocal tract noise response, $r(t)$. The **STV** coarticulation channel then models specifically the transformation of the vocal tract from its /V/ articulation [$r1(t)$] to its C/V/C articulation [$r2(t)$].

Finally, let the $P(a,b)$ characterization associated with this coarticulation channel be called **COAR**(s,τ). The function **COAR**(s,τ) is composed of the same wavelet-

coefficient distribution as $P_{SE}(a,b)$. The **COAR**(s,τ) is merely a special case of the $P_{SE}(a,b)$, as defined above. In this case, the scale and shift parameters (s,τ) are used in place of (a,b).

The P_{SE} coarticulation model is thus defined as follows. The initial (control) state is specified by one articulation of the vocal tract, /V/. The effected state is manifested (relative to the first) by a second articulation of the vocal tract, C/V/C. The noise response function $[r(t)]$ serves the purpose of characterizing each of these vocal tract articulations. The transformational mapping from one speech state to the other is depicted in **COAR**(s,τ).

The **COAR** model expressed in analytic terms is analogous to the P_{SE} model of equation [4.1]:

$$\textbf{[4.4]} \qquad \mathbf{STV}_{\mathbf{COAR}(s,\tau)}[r1(t)] = r2(t)$$

where:

- $r1(t)$ is the vocal tract noise response corresponding to the isolated vowel articulation, /V/.
- $r2(t)$ is the vocal tract noise response corresponding to the contextual vowel articulation, C/V/C.
- s,τ are the time-scale and time-shift wavelet parameters. (These have been used in place of a,b.)
- **COAR**(s,τ) is the **STV** channel characterization of the coarticulation effect. It describes the transformation from the isolated vowel articulation to the contextual vowel articulation.

The speech coarticulation model of [4.4] is illustrated in Figure 4.2 below:

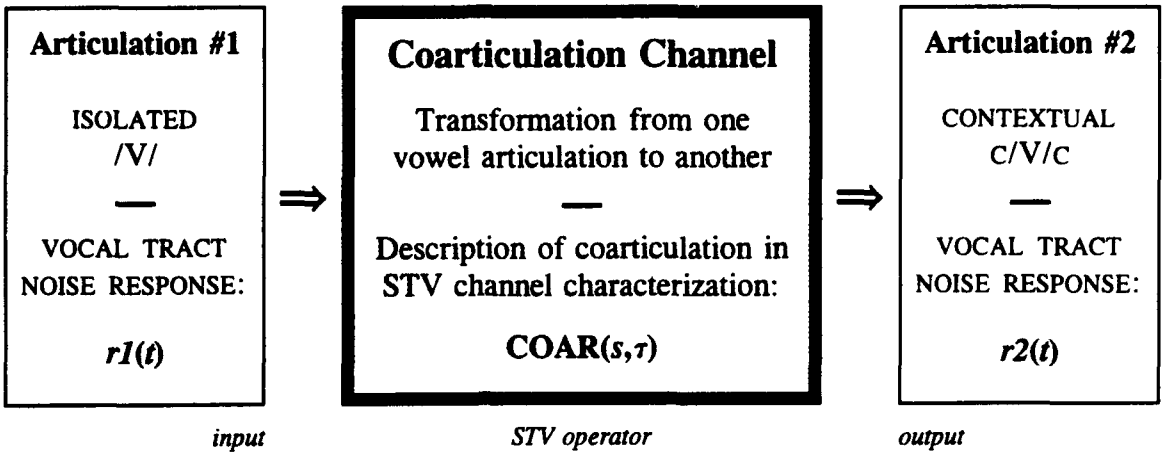


Figure 4.2 **The Vocal Tract Coarticulation Channel**

As depicted in the figure, the **COAR**(s,τ) distribution specifically indicates (with respect to each vowel formant) the time-shift intervals and scale values which differentiate *r2* from *r1*, thereby providing a description of the overall coarticulation effect. This description, like any wavelet transform, is three-dimensional:

magnitude of the correlation (for each shifted/scaled component)

vs. *time-shift interval* (time-dependence of the transition)

vs. *scale value* (scale factor for the transition).

As noise response functions, *r1* and *r2* describe the time-varying behavior of the vocal tract. The **COAR**(s,τ), therefore, also specifically models the behavior of the vocal tract. As such, the **COAR**(s,τ) gives a time-varying description of speech *articulation*, as opposed to a one of speech *production*.

It should be further noted that the **COAR**(s,τ) is an instance of the more general speech-effect representation, $P_{SE}(a,b)$. The **COAR** system may be considered as one realization of an P_{SE} speech effect. The **COAR**, restricted to the domain of the vocal tract, specifically addresses consonant-vowel-consonant coarticulation.

4.6 Estimating the COAR

The coarticulation channel is characterized by the function **COAR**(s,τ). As is the case for $P_{SE}(a,b)$, the **COAR**(s,τ) distribution can be estimated by the wavelet transform of the system output taken with respect to the system input. As an analogy to equation [4.2], therefore, the estimate of the **COAR**(s,τ) appears:

$$\textbf{[4.5]} \qquad \widehat{\textbf{[COAR]}}\ (s,\tau) \;=\; \mathbf{W}_{r1(t)}\, r2(t) \quad (s,\tau)$$

where $\widehat{\textbf{[COAR]}}$ denotes an estimate of **COAR**.

Unlike the input and output signals of the P_{SE} system, the signals *r1* and *r2* are strictly defined. Ideally, the only difference between the utterances from which *r1* and *r2* are generated is the presence/absence of segmental coarticulation. As a result, the estimate $\widehat{\textbf{[COAR]}}(s,\tau)$ manifests precisely the coarticulation effect of /C−C/ on /V/. $\widehat{\textbf{[COAR]}}(s,\tau)$ so describes the CVC coarticulation attributable to a particular vowel-consonant combination produced by one speaker.

4.7 Model Summary

The construct $P_{SE}(a,b)$ is the focus of a proposed wavelet model for the analysis of speech effects. The $P_{SE}(a,b)$ functions in conjunction with an **STV** operator (for waveforms) which executes a speech-effect transformation. Figure 4.1 depicts the general structure of the P_{SE} model.

As a special case of the P_{SE}, the **COAR** system models vocal tract *articulation*. The channel **COAR**(s,τ) operates on vocal tract noise response functions. The input/output utterances associated with these functions are defined specifically for the purposes of highlighting the effects of CVC coarticulation. As a result, the variances which might be generated from *other* components of speech production (such as the laryngeal source) are, primarily, eliminated from the model. The structure of the **COAR** model appears in Figure 4.2.

Both constructs [$P_{SE}(a,b)$ and **COAR**(s,τ)] are wavelet-domain functions, and they are practically estimated by the appropriate wavelet transform. Inherent to the structure of these models is the formulation of a contrast between two separate utterances. In addition to their primary role as speech-effect descriptors, it is shown that the $P_{SE}(a,b)$ and **COAR**(s,τ) may be viewed as:

1) comparative *correlations* between the effected state and control state,

or 2) recipes for *generating* the effected state from the control state.

Chapter 5

SOLUTION

In the previous section, the speech model is presented in abstract terms. What remains is to evaluate the model using samples of real speech. For the purposes of implementation and evaluation, this study focuses on the particular case of the coarticulation problem. The **STV** wavelet model for vocalic speech coarticulation appears in Figure 4.2. From the point-of-view of describing a coarticulatory effect, the "problem" is identifying the model's **STV** channel characterization. The function $\mathbf{COAR}(s,\tau)$ thus appears as the unknown quantity. A practical method for measuring the $\mathbf{[C\hat{O}AR]}(s,\tau)$ (associated with a particular vowel-consonant combination and a particular speaker) is therefore desired.

The formulation of the $\mathbf{COAR}(s,\tau)$ which appears in the previous section (equation [4.4]) is stated in terms of vocal tract noise response functions $[r(t)]$. However, the noise response for a real vocal tract cannot be measured directly. The purpose of this section, therefore, is to formulate the $\mathbf{[C\hat{O}AR]}(s,\tau)$ in terms of directly measurable quantities. A solution for $\mathbf{[C\hat{O}AR]}(s,\tau)$ as a function of $z(t)$, the voice output response function, is presented here.

The analysis which leads to this solution includes a number of steps. First, the $\mathbf{[C\hat{O}AR]}(s,\tau)$ is recognized as the wavelet transform $\mathbf{W}_{r1}r2$. Each step which follows obtains a progressively more-concrete version of $\mathbf{W}_{r1}r2$. The final form of $\mathbf{W}_{r1}r2$ is a function of: $z1(t)$, the voice output response associated with the isolated /V/ utterance,

and $z2(t)$, the voice output response associated with the contextual C/V/C utterance. This form thus serves as the measurable estimate $\mathbf{[C\hat{O}AR]}(s,\tau)$, suitable for evaluation on real speech utterances.

5.1 COAR via the Mother Mapper

Equations [3.8] and [3.9] show how the mother mapper may serve as a critical tool for calculating certain wavelet transforms. The objective of this analysis, $\mathbf{[C\hat{O}AR]}(s,\tau)$, is one of those wavelet transforms.

Let $\mathbf{W}_{r1}r2$ be expressed as a function of two other wavelet transforms. Let $n(t)$ serve as the common wavelet for each of these functional wavelet transforms. Substituting these functions into equation [3.8] yields:

$$\mathbf{W}_{r1}r2 \;=\; \text{an integral function of} \quad \mathbf{W}_{n}r2 \quad \text{and} \quad \mathbf{W}_{n}r1$$

Though the noise function $n(t)$ hardly qualifies as "analytic," this choice of substitution for $f(t)$ serves the present purposes. Note that $\mathbf{W}_{r1}r2$ appears in equation [4.5] as the $\mathbf{[C\hat{O}AR]}(s,\tau)$. Using the mother mapper integral in equation [3.9], the above relation appears explicitly as:

$$\mathbf{[C\hat{O}AR]}(s,\tau) \;=\; \mathbf{W}_{r1}r2\,(s,\tau)$$

$$\mathbf{[5.1]} \qquad = \frac{1}{C_n}\int\frac{1}{a^2}\int \mathbf{W}_{n}r2\,(a,b)\cdot\mathbf{W}^{*}_{n}r1\left(\frac{a}{s},\frac{b-\tau}{s}\right)db\,da$$

where C_n is the normalizing constant for $n(t)$. Equation [5.1] is thus an estimate for the **COAR**(s,τ) expressed in terms of two "new" wavelet transforms. The first is the wavelet transform of the vocal tract noise response associated with the C/V/C articulation. The second is the wavelet transform of the vocal tract noise response associated with the /V/ articulation. Both wavelet transforms are taken with respect to the same broadband noise function, $n(t)$. $\mathbf{W}_n r1$ is scaled and shifted with respect to $\mathbf{W}_n r2$.

5.2 The COAR Estimate in Abstract Form

In the previous section, the mother mapper is used to derive an estimate for the **COAR**(s,τ) which is stated in terms of two wavelet transforms. In this section is added the principles of the **STV** operator formulation.

Equivalent expressions for:

$$\mathbf{W}_n r2\ (a,b) = \hat{P2}\ (a,b) \qquad \text{and} \qquad \mathbf{W}_n r1\ (a,b) = \hat{P1}\ (a,b)$$

which are found in equation [3.6], are substituted into the integral equation [5.1]:

$$[\hat{\mathbf{COAR}}]\ (s,\tau) = [\hat{\mathbf{W}}]_{r1} r2\ (s,\tau)$$

$$= \frac{1}{C_n} \int \frac{1}{a^2} \int \hat{P2}\ (a,b) \cdot \hat{P1}^* \left(\frac{a}{s}, \frac{b-\tau}{s} \right) db\ da \qquad \textbf{[5.2]}$$

where:

$[\hat{\mathbf{W}}]$ denotes an estimate of the wavelet transform coefficient **W**.

$\hat{P1}^*$ denotes the complex conjugate operation on the **STV** channel characterization estimate $\hat{P1}$.

$\hat{P1}\,(^{a}/_{s}\,,\,^{b-\tau}/_{s})$ denotes the scaling ($^{1}/_{s}$) and shifting ($b-\tau$) of the **STV** channel characterization estimate $\hat{P1}$.

Equation [5.2] is an expression for the estimate **[CÔAR]**(s,τ) stated in terms of estimates for $P1(a,b)$ and $P2(a,b)$. The functions $P1$ and $P2$ are the **STV** channel characterizations (of the *vocal tract* channel) associated with each of the two articulations /V/ and C/V/C. One of the channel representations, $\hat{P1}(a,b)$, is scaled and shifted with respect to the other, $\hat{P2}(a,b)$. This scaling and shifting in a,b is effected through the analogous parameters s,τ. Integration occurs over the arguments a,b, of which $\hat{P1}$ and $\hat{P2}$ are functions. The overall expression is a function of the shifted/scaled parameters s,τ.

As previously stated, the functions $P1(a,b)$ and $P2(a,b)$ are wavelet descriptions of the vocal-tract channel. Equation [5.2], therefore, is an expression for the **[CÔAR]**(s,τ) in which all of the components (except for the constant C_n) are independent of any excitation or source function.

5.3 The COAR Estimate in Realizable Form

Equation [5.2] remains a theoretical relation in terms of an abstract representation of the vocal tract $[P(a,b)]$. In order to measure the **[CÔAR]**(s,τ) practically, the

estimates $\hat{P}1(a,b)$ and $\hat{P}2(a,b)$ must be evaluated. For each, a realizable form can be found, namely, the one derived from a *real* excitation [$g(t)$] and a real voice output response [$z(t)$]. From equation [3.6]:

$$\hat{P2}\ (a,b) = \mathbf{W}_{g2} z2\ (a,b) \qquad \text{and} \qquad \hat{P1}\ (a,b) = \mathbf{W}_{g1} z1\ (a,b)$$

Substituting into equation [5.2]:

$$[\hat{\mathbf{COAR}}]\,(s,\tau) = [\hat{\mathbf{W}}]_{r1} r2\ (s,\tau)$$

$$= \frac{1}{C_n} \int \frac{1}{a^2} \int \mathbf{W}_{g2(t)} z2 \cdot \mathbf{W}^{*}_{g1(t)} z1 \left(\frac{a}{s}, \frac{b-\tau}{s}\right) db\ da \qquad \textbf{[5.3]}$$

The wavelet transform $\mathbf{W}_{g2(t)}z2$ is a function of a,b. $\mathbf{W}_{g1(t)}z1$ is scaled and shifted by s,τ.

Equation [5.3] thus gives an estimate for the **COAR**(s,τ) expressed in terms of potentially measurable parameters. The wavelet transforms are effected for $z1$ and $z2$ (which are the voice output responses derived from two real, complete utterances), using the mother wavelets $g1$ and $g2$ (which are the glottal-source time functions for these utterances).

5.4 Determination of the Glottal Source Function

If the result in equation [5.3] is to be utilized, then some method for measuring and/or approximating the glottal source function [$g(t)$] for each of two utterances is yet required. Following is an explanation of three potential solutions to this problem.

The first solution is to derive $g(t)$ using a sampled version of the signal at the microphone [$z(t)$]. Some of the signal processing techniques available for separating the glottal function from the vocal tract impulse response function may then be employed. These methods (such as the cepstral filtering technique) generally assume a stationary (time-invariant) model of the vocal tract impulse response (Saito and Nakata 1985). A model of this type is inadequate for the present purposes. (An assumption of stationarity for the *glottal* function [$g(t)$], however, over the course of a single isolated vowel or C/V/C utterance, might be reasonable.)

This solution to finding $g(t)$ poses a further problem of interpolating between time-samples. Interpolation arises because the samples of $g(t)$ would constitute a discrete and random time-series. The family of *scaled* versions of $g(t)$ is a necessary ingredient for the expression in [5.3]. These scaled versions of $g(t)$ appear as wavelets in the wavelet transforms $\mathbf{W}_{g2}z2$ and $\mathbf{W}_{g1}z1$, and they take the same form as does the function f in equation [3.2]. In order to derive these scaled versions of $g(t)$, a knowledge of the time-series at intermediate sample-times is required.

One method for avoiding the interpolation problem is to reformulate the wavelet transforms appearing in equation [5.3]. Each can be expressed in terms of two other

wavelet transforms, using the mother mapper of equation [3.8]. Such a reformulation would employ the use of a standard analyzing wavelet [$f(t)$] which is specified analytically. In the resulting expression, that function would serve as the mother wavelet in four wavelet transforms: $\mathbf{W}z2$, $\mathbf{W}g2$, $\mathbf{W}z1$, and $\mathbf{W}g1$.

The second potential method of finding $g2$ and $g1$ for equation [5.3] is to approximate the functions analytically. In a vocalic utterance, the shape of the periodic glottal function [$g(t)$] is primarily influenced by two speech parameters, fundamental frequency (*F0*) and voice intensity (Miller 1959). For the purposes of deriving an analytic approximation to $g(t)$, each of these parameters could be controlled by the speaker and/or measured directly. How effectively such an approximation would serve the estimate to $\mathbf{COAR}(s,\tau)$ is not known. Some investigation would be necessary in order to determine how dramatically the errors in the $g(t)$ approximation would propagate through computation of the wavelet transforms $\mathbf{W}_{g2}z2$ and $\mathbf{W}_{g1}z1$.

The third approach to specifying the glottal functions $g2$ and $g1$ is contingent on an assumption. Assume:

[5.4]
$$g2(t) = g1(t)$$

Implicit in this assumption are the following necessary (but probably not sufficient) conditions (Flanagan and Cherry 1969; Monsen and Engebretson 1977; Rothenberg 1973):

1) The two utterances, /V/ and C/V/C, are produced by the same speaker.

2) The *F0* (fundamental frequency) of each utterance is constant across the utterance.

3) The *F0* of utterance #1 equals the *F0* of utterance #2.

4) Throughout each utterance, the intensity of $g(t)$ is a constant.

5) The intensity of $g(t)$ for utterance #1 equals the intensity of $g(t)$ for utterance #2.

If the assumption in [5.4] can be afforded, then equation [5.3] becomes:

$$[\hat{\mathbf{COAR}}](s,\tau) = [\hat{\mathbf{W}}]_{r1}\, r2\,(s,\tau)$$

$$= \frac{1}{C_n}\int \frac{1}{a^2}\int \mathbf{W}_{g2(t)}z2 \cdot \mathbf{W}^{*}_{g2(t)}z1\left(\frac{a}{s}, \frac{b-\tau}{s}\right) db\, da \qquad [5.5]$$

The wavelet transform $\mathbf{W}_{g2(t)}z2$ is a function of a,b. $\mathbf{W}_{g2(t)}z1$ is scaled in s and time-shifted in τ.

5.5 The COAR Estimate in Measurable Form

The final form of the $[\hat{\mathbf{COAR}}](s,\tau)$ is derived from the glottal-source condition in [5.4] and a formulation of the mother mapper.

First, the mother mapper is used to re-express the wavelet transform of $z2(t)$ with respect to $z1(t)$. Specifically, $\mathbf{W}_{z1}z2$ is stated in terms of $\mathbf{W}_{g2}z2$ and $\mathbf{W}_{g2}z1$. In other words, using equation [3.9], $\mathbf{W}_{z1}z2$ appears in place of $\mathbf{W}_{y}x$, and $g2(t)$ appears in place of f. These substitutions result in the following expression:

$$\mathbf{W}_{z1(t)}z2(t)\ (s,\tau) \;=\; \frac{1}{C_{g2}}\int \frac{1}{a^2}\int \mathbf{W}_{g2(t)}z2\ (a,b)\cdot \mathbf{W}^{*}_{g2(t)}z1\left(\frac{a}{s},\frac{b-\tau}{s}\right)db\,da$$

Combining the above equation with equation [5.5] yields:

$$(C_n)\cdot[\hat{\mathbf{COAR}}]\,(s,\tau) \;=\; (C_{g2})\cdot \mathbf{W}_{z1(t)}z2(t)\ (s,\tau)$$

So that:

[5.6] $$\mathbf{W}_{z1(t)}z2(t)\ (s,\tau) \;=\; \left(\frac{C_n}{C_{g2}}\right)\cdot[\hat{\mathbf{COAR}}]\,(s,\tau)$$

The mother mapper is utilized once more to reformulate (for the second time) the wavelet transform of $z2(t)$ with respect to $z1(t)$. This time, however, $\mathbf{W}_{z1}z2$ is re-expressed in terms of $\mathbf{W}_f z2$ and $\mathbf{W}_f z1$. $f(t)$ is an analytic mother wavelet function. Using equation [3.9], $\mathbf{W}_{z1}z2$ appears in place of $\mathbf{W}_y x$:

$$\mathbf{W}_{z1(t)}z2(t)\ (s,\tau) \;=\; \frac{1}{C_f}\int \frac{1}{a^2}\int \mathbf{W}_{f(t)}z2\ (a,b)\cdot \mathbf{W}^{*}_{f(t)}z1\left(\frac{a}{s},\frac{b-\tau}{s}\right)db\ da$$

Combining the above equation with equation [5.6] gives:

$$\left(\frac{C_n}{C_{g2}}\right)\cdot[\hat{\mathbf{COAR}}]\,(s,\tau) \;=\; \frac{1}{C_f}\int \frac{1}{a^2}\int \mathbf{W}_{f(t)}z2\cdot \mathbf{W}^{*}_{f(t)}z1\left(\frac{a}{s},\frac{b-\tau}{s}\right)db\ da$$

Finally, the constants C_n, C_{g2}, and C_f are combined into one constant C_x:

$$C_x \equiv \left(\frac{C_n \cdot C_f}{C_{g2}} \right)$$

And:

[5.7] $$[\hat{\mathbf{COAR}}](s,\tau) = \frac{1}{C_x} \int \frac{1}{a^2} \int \mathbf{W}_{f(t)} z2 \cdot \mathbf{W}^*_{f(t)} z1 \left(\frac{a}{s}, \frac{b-\tau}{s} \right) db\ da$$

where:

COAR(s,τ) is a wavelet description of the coarticulation in terms of an **STV** channel characterization. The "channel" transforms an isolated /V/ into a coarticulated C/V/C (initial and final consonants the same). This distribution is specified for a particular vowel-consonant combination as produced by one speaker.

s,τ are the wavelet parameters time-scale and time-shift.

$z(t)$ is the voice output response measured at the microphone when the vocal tract is excited by a real glottal source $g(t)$.

$z2$ is the voice output response signal associated with the contextual C/V/C articulation.

$z1$ is the voice output response signal associated with the isolated /V/ articulation.

$f(t)$ is a standard analyzing mother wavelet, known analytically.

And the following assumption must be satisfied:

[5.4] $$g2(t) = g1(t)$$

Thus, given the assumption that the glottal source time-functions are equal, equation [5.7] provides an estimate for the **COAR**(s,τ) expressed exclusively in terms of *measurable* quantities. $z1(t)$ and $z2(t)$ can be recorded as the voltage output from a standard speech microphone. $f(t)$ is known in the form of an analytic expression. In practice, therefore, it is necessary to ensure some measure of uniformity between the voicing/excitation functions associated with test utterances *z1* and *z2*.

This point brings up the question of what happens when the mathematical assumption in equation [5.4] is *not* satisfied. Because equation [5.7] is expressed in terms of the voice output responses of the test utterances [$z1(t)$ and $z2(t)$], the estimate **[CÔAR]**(s,τ) effectively measures the transformation between these (complete) utterances. To the extent that utterances $z1(t)$ and $z2(t)$ have similar voicing conditions (in a phonological sense), then the **[CÔAR]**(s,τ) estimate indeed differentiates between the *vocal tract articulatory states* associated with these utterances.

In other words, in the case that equation [5.4] exactly holds, the **[CÔAR]**(s,τ) contrasts between the vocal tract states of the test utterances, as claimed in the original model (Figure 4.2). However, the form of the expression in equation [5.7] yields a contrast between the $z1(t)$ and $z2(t)$ voice output responses. Therefore, it is concluded that whenever the phonological voicing conditions in these utterances are similar, the **[CÔAR]**(s,τ) estimate yet measures the contrast between vocal tract articulatory states.

In short, the assumption in equation [5.4] is interpreted as an assumption of uniformity in voicing. Under uniform conditions of voicing, the isolated /V/ and contextual C/V/C utterances differ primarily in aspects related to the absence or presence

of coarticulation. These qualities, reflected in $z1(t)$ and $z2(t)$, are set into contrast by the **[CÔAR]**(s,τ) in its measurable form (equation [5.7]). The **[CÔAR]**(s,τ) is thus rendered as a description of the coarticulation present in C/V/C relative to /V/.

Chapter 6

EXPERIMENT

6.1 A Study to Evaluate the Model

The proposed model for speech coarticulation was evaluated experimentally, using samples of human utterances. In particular, the **COAR**(s,τ) was formulated between pairs of real articulations and calculated using the method of estimation given in the previous section. Each articulation pair consists of one isolated vowel and the same vowel appearing in a C/V/C context. The **COAR**(s,τ) function thereby depicts the transformation of the vowel from the isolated case to the contextual case.

The initial and final consonants in CVC are always the same. The estimate **[CÔAR]**(s,τ), therefore, generates a description of C/V/C coarticulation for that vowel-consonant combination. One speech subject produced all of the utterances in the experiment.

A series of **[CÔAR]**(s,τ) estimates were so calculated for a variety of articulations. 4 different vowels were examined in the company of 7 different consonants. The speech sample thus consists of 28 different CVC combinations. Furthermore, 4 repetitions of these 28 combinations were included. The consonant sample includes stops, nasals, and liquids.

The model was evaluated on the basis of how effectively the **COAR**(s,τ) description reflects these phonemic variations. For example:

1) How does the appearance of the **COAR**(s,τ) distribution change for different vowels used in the pair?

2) How does the **COAR**(s,τ) distribution change when different consonants are used for context?

3) Do changes in the **COAR**(s,τ) distribution appear to correlate with vowel place-of-articulation? Do they ap[illegible]late with consonantal place-of-articulation?

4) Does the **COAR**(s,τ) illuminate vocalic nasality?

In short, do parameters of the **COAR**(s,τ) distribution correlate with *phonetic* parameters, such as place and manner of articulation? The purpose of the experiment was to provide evidence for answering these questions.

The *goal* of the experiment was to determine whether the dimensionality of the coarticulation problem is effectively lowered by the introduction of this coarticulation model. The value of the model is contingent on whether it can acoustically parameterize phonetic variables in a concise manner. A concise description of CVC coarticulation, one which is applicable to a variety of phonemic contexts, would contribute to our understanding of continuous speech, both for the purposes of its clinical production and its synthetic generation.

6.2 Implementation of the COAR Solution

The measurable form of the **COAR**(s,τ) estimate is:

$$[5.7] \qquad \widehat{[\mathbf{COAR}]}(s,\tau) = \frac{1}{C_x}\int\frac{1}{a^2}\int \mathbf{W}_{f(t)}\, z2 \cdot \mathbf{W}^{*}_{f(t)}\, z1\left(\frac{a}{s}, \frac{b-\tau}{s}\right) db\, da$$

$z1(t)$ is the recorded microphone signal of the isolated /V/ utterance, and $z2(t)$ is the recorded signal of the contextual C/V/C. Recall that this estimate for the **COAR**(s,τ) calls for the following assumption to be satisfied:

$$[5.4] \qquad g2(t) = g1(t)$$

where $g1(t)$ and $g2(t)$ are the glottal source time-functions associated with each utterance. As stated previously (page 56), equation [5.4] is interpreted as an assumption of uniformity in voicing between the utterances $z1(t)$ and $z2(t)$.

The speech subject was therefore trained to produce pairs of utterances in $z1$ and $z2$ which met the following qualifications:

1) The fundamental frequency over the vocalic portion of each utterance was constant throughout the utterance and equal within the pair.

2) The intensity over the vocalic portion of each utterance was constant throughout the utterance and equal within the pair.

The criteria for constant fundamental frequency was that the utterances of a pair were deliberately sustained with constant pitch, and that they were perceived, by the subject and experimenter alike, to exhibit constant pitch. The experimenter's good proficiency

in music substantiates his capacity for perceiving voice pitch in this respect. The criteria for constant intensity was that the utterances of a pair were deliberately sustained with the same vocal effort, and that they were perceived by the experimenter to have the same loudness.

The function $f(t)$ is an analyzing mother wavelet. For all of the wavelet transforms implemented in this study, the choice of mother wavelet was the Morlet ($\omega_0 = 41.77\ ms^{-1}$). The criteria for this selection is stated in the appendix (page 161).

6.3 The Speech Sample

The speech sample word list includes the vowel and consonant phones which appear in Table 6.1 (Ladefoged 1975; Stevens and House 1963, p. 114):

Table 6.1 **The Speech Sample Phones**

Vowels /V/		**Consonants /C/**			
/i/ beat	/u/ boot	***stops***	/b/ by	/d/ dye	/g/ guy
/æ/ bat	/ä/ father	***nasals***	/m/ my	/n/ nigh	
		liquids	/r/ rye	/l/ lie	

Each C/V/C utterance from the speech sample begins with one of these seven consonants and finishes with the same consonant. One of the four vowels appears between them. The same vowel sustained alone, /V/, constitutes the isolated utterance of the pair.

With respect to the voiced/unvoiced distinction of the consonant appearing in the C/V/C utterances, it is expected that the vowel might be subject to some variability (Stevens and House 1963, p. 121). The variables of speech production which are of interest in this study, however, are limited to those associated with articulations of the vocal tract, i.e., to the variables place of articulation and manner of articulation. For this reason, the voiced/unvoiced distinction does not appear in the consonant list; only voiced consonants were examined.

In his study on spectrographic vowel reduction, Lindblom (1963) showed that the duration of a /CVC/ syllable determines the degree to which a vowel undergoes contextual modification or "reduction". A longer vowel duration tends to facilitate the vowel articulation reaching its "target". A shorter vowel duration, on the other hand, provides less time for the articulators to complete their glide movements from /C/ to /V/ and back to /C/ again. The result is a *more reduced* /V/ in the case of the shorter vowel. In the current regard, therefore, **COAR**(s,τ), which describes CVC coarticulation, is also expected to be a function of the C/V/C vowel's duration.

For the following reason, however, the speech sample in this study does *not* include duration as a variable factor. The "vowel reduction" effect attributable to coarticulation is the result of the articulators moving at finite velocities to and from their sequential phonemic targets (Lindblom 1963, pp. 1778-1779). The phys ponse of

a given articulator, therefore, is smooth and continuous. The overall shape of the vocal tract is a continuous and dynamic function of time. The *longest* duration C/V/C vowel is thus expected to include an entire range of "reduced" articulations. They begin from the most reduced (most modified) version which immediately follows the initial /C/, they vary continuously to the target /V/, and they evolve back to the reduced version in anticipation of the final /C/. In theory, the long duration C/V/C vowel includes in its subset any reduced version generated from the short duration C/V/C.

The current speech sample is thus composed deliberately from "long" C/V/C vowels. By this it is meant that each utterance is treated as a complete isolated word; yet, no utterance was sustained for an unnaturally long duration. In particular, the C/V/C was spoken alone, produced with stress, and void of any semantic context. The isolated /V/ of the pair, which represents the steady-state "target" articulation, was likewise spoken alone with stress. The durations of all vowels, appearing either alone or in C/V/C, were roughly constant. However, because the wavelet transform W_{z1z2} requires no such restriction, no special effort was made to ensure *exact* equality between the durations of vowels within a /V/, C/V/C pair.

Table 6.2 shows all of the 28 utterances included in the word list.

6.4 The Speech Subject

One male subject was used for the production of all utterances in the word list. He is a native American English speaker, age 27. His speech was assessed by a speech

Table 6.2 **The Word List**

KEY:

isolated /V/
contextual C/V/C
phonetic spelling

STOPS			NASALS		LIQUIDS	
bilabial	**alveolar**	**velar**	**bilabial**	**alveolar**	**retro**	**alveolar**
1 /i/ /bib/ beeb	10 /i/ /did/ deed	19 /i/ /gig/ geeg	28 /i/ /mim/ meem	37 /i/ /nin/ neen	46 /i/ /rir/ rear	55 /i/ /lil/ leel
4 /æ/ /bæb/ babb	13 /æ/ /dæd/ dad	22 /æ/ /gæg/ gag	31 /æ/ /mæm/ ma'am	40 /æ/ /næn/ nan	49 /æ/ /rær/ rær	58 /æ/ /læl/ lal
6 /ä/ /bäb/ bob	15 /ä/ /däd/ dodd	24 /ä/ /gäg/ gogg	33 /ä/ /mäm/ mom	42 /ä/ /nän/ non	51 /ä/ /rär/ raar	60 /ä/ /läl/ laal
9 /u/ /bub/ boob	18 /u/ /dud/ dude	27 /u/ /gug/ goog	36 /u/ /mum/ moom	45 /u/ /nun/ noon	54 /u/ /rur/ rure	63 /u/ /lul/ lool

and language pathologist to be standard American, having a general American dialect and no articulation errors. His hearing was assessed by an audiologist as normal. In particular, the subject measured to within +5 dB hearing level at pure tone frequencies from 250 Hz to 8 kHz, and he performed satisfactorily on a series of word discrimination tests.

The use of the speaker as a human research subject in this capacity was reviewed and approved by the Human Subjects Institutional Review Board of The Pennsylvania State University on January 22, 1993.

6.5 Instructions to the Subject

The **COAR**(*a*,*b*) was always calculated from utterances produced in *pairs* [/V/, C/V/C]. The objective of this design is to maximize the likelihood that the differences between the isolated and CVC versions of the vowel were primarily attributable to segmental coarticulation. Therefore, the isolated /V/ utterance always immediately preceded the C/V/C utterance in the pair.

The subject was familiarized with correct pronunciation of the utterances in the word list through the audition of numerous examples. He received the following visual cue for each utterance pair:

example pair 1: Say **/i/** as in *beat*.
Say **beeb**.

example pair 2: Say **/u/** as in *boot*.
Say **noon**.

For the purposes of maintaining allophonic consistency among the stop-consonant articulations, the subject was instructed to produce (in final position) an *exploded* stop.

After familiarization with correct pronunciation, the subject was instructed to maintain a constant loudness from one utterance to the next within each pair. He was also instructed to maintain a constant pitch within each pair. To help the subject execute constant fundamental frequency, correct and incorrect examples of constant-pitch utterances were played for the subject to hear before recording. To help the subject execute constant intensity during recording, a visual monitor was provided by means of a calibrated VU intensity level meter.

The speech subject produced 4 repetitions of each utterance pair in the word list. The repetitions were generated from 4 individually randomized sets, each consisting of 28 utterance pairs. Every utterance pair in the word list occurs once within the random set.

6.6 Processing the Speech Signal

The speaker's utterances were recorded in a sound treated room using an Electro-Voice RE20 dynamic cardioid microphone and a SONY TCD-D10 digital audio tape recorder. The analog output from the tape recorder was low-pass filtered at 8 kHz using an 8-pole Butterworth filter.

To prevent aliasing in the transition-band of the filter, the signal was over-sampled at a frequency of 31.25 kHz. The analog-to-digital conversion was performed

by an ARIEL DSP-16 processing board hosted on an AT&T PC 6300 computer. The binary time-series obtained from the digital signal processing board was uploaded to an IBM RISC 6000 workstation, which was used to perform all of the wavelet calculations. The time-series data were also recorded onto an optical archival-quality medium. Software for the implementation of the wavelet transform and mother mapper integrals was written specially for this application in the Ada programming language.

6.7 The Wavelet Transform Grid Spacing

This section specifies the range, number, and density of discrete wavelet-coefficient "points" evaluated in the (a,b) domain for all of the wavelet transforms calculated in this study. Table 6.3 states the explicit functions used for wavelet transform evaluation and their parameterization in scale a and shift b:

Table 6.3 **Morlet Wavelet** $f_M(t)$

The Morlet mother wavelet: $$f_M(t) = e^{j\,(41.77)\,t} \cdot e^{\left(-\frac{t^2}{2}\right)}$$

The Morlet wavelet family: $$f_M(t)_{a,b} = e^{j\,(41.77)\,\left(\frac{t-b}{a}\right)} \cdot e^{\left[-\frac{\left(\frac{t-b}{a}\right)^2}{2}\right]}$$

$$f_M(t)_{1,0} = f_M(t)$$

The function $f_M(t)_{1,0}$ corresponds to a Gaussian-windowed complex-exponential which is centered at the spectral frequency 6.648 kHz. This function has a 3 dB time-window width equal to 1.7 milliseconds and a 3 dB spectral bandwidth equal to 265 Hz.

The scale and shift parameters are evaluated numerically on a discrete grid. Table 6.4 states the range and interval for the scale parameter evaluations:

Table 6.4 **Scale Factor** a

minimum a = 1.662	associated spectral frequency = 4.00 kHz	= $f_{MAXIMUM}$
maximum a = 60.23	associated spectral frequency = 110 Hz	= $f_{MINIMUM}$
a is incremented by a multiplicative factor **1.0369322** for each consecutive evaluation		
The total number of frequency points on the grid is **100**		

Thus, the range of frequencies evaluated by the wavelet transform extends from 110 Hz (just *higher* than the subject's fundamental frequency) to 4 kHz (well above third formant frequency for any of the subject's vowels). The scale evaluations are geometrically spaced; i.e., they are separated by a constant multiplicative factor. This means that the logarithm of a is evenly spaced by the interval log(1.0369322) = 0.0158 log scale units.

The reason for evaluating the wavelet transform only at frequencies higher than the fundamental is that vowel behavior is ultimately manifested in the formant structure. Furthermore, the fundamental frequency of excitation was controlled by this experiment in a manner designed to remove it as a potential source of variation in the **COAR**(a,b). The *benefit* derived from excluding the fundamental frequency from the wavelet transform grid is a more limited operating range (110 to 4,000 Hz). A more limited scale/frequency range yields in the analysis wavelet more optimal simultaneous resolution in time and frequency.

Table 6.5 states the range and interval for evaluating the time-shift parameter in these wavelet transforms:

Table 6.5 **Time-Shift Parameter** b

b is incremented every **2.496** milliseconds throughout the entire duration of the utterance
The total number of time points on the grid is **400**

More specific information about the (a,b) grid spacing, including the effective resolution bandwidths associated with each time-frequency "bin," appears in Table 6.6.

The complete grid spacing in a, (i.e., the incremental factor combined with the frequency bandwidths shown above) results in an overlap between adjacent frequency bins. The amount of overlap (using the 3 dB criterion) is roughly 7%.

Table 6.6 **The Wavelet Transform Scale Grid**

Scale-Point Number	Scale Factor	Associated Spectral Frequency	3 dB Frequency Bandwidth	3 dB Time-Window Width
	$a = 1.0$	$f_{1.0} = 6.6$ kHz	$BW_{1.0} = 265$ Hz	$TW_{1.0} = 1.7$ ms
N = 99	$a_{MIN} = 1.7$	$f_{MAX} = 4.0$ kHz	$BW_{MAX} = 159$ Hz	$TW_{MIN} = 2.8$ ms
N = 64	$a = 5.9$	f = 1.1 kHz	BW = 45 Hz	TW = 9.8 ms
N = 50	$a = 9.8$	f = 677 Hz	BW = 27 Hz	TW = 16 ms
N = 35	$a = 17$	f = 393 Hz	BW = 16 Hz	TW = 28 ms
N = 0	$a_{MAX} = 60$	$f_{MIN} = 110$ Hz	$BW_{MIN} = 4.4$ Hz	$TW_{MAX} = 100$ ms

The grid spacing in b also results in an overlap between adjacent time-windows. Considerably more time-window overlap occurs at the maximum scale value than at the minimum scale value. In the minimum case, however, $a = 1.7$ yields a 3 dB time-window width of 2.8 milliseconds. With b incremented every 2.5 milliseconds, the amount of overlap between adjacent time-windows in this region is roughly 10%.

6.8 The Relationship Between *z2* and C/V/C

The coarticulation model as outlined in the previous sections consists of a correlation or "cross-wavelet" formulated between the signal recorded from an isolated vowel and that from another *vowel* appearing in a C/−/C context. This would suggest that, for the purposes of implementation, the signal associated with the vowel portion of the CVC utterance must be "segmented-out" or removed from its context. Although, in many circumstances, a segmentation between vowel and consonant would constitute a standard waveform-editing procedure. such a segmentation was *not* employed in this study.

The reason for avoiding a segmentation of the V from C/−/C is that it assumes the consecutive phones [C, V, C] will be manifested discretely in-sequence within the acoustic domain of the signal waveform. This assumption violates the basis of a continually time-variant model for CVC coarticulation, whereby, the articulatory effects of the initial and final consonant are manifested *throughout* portions of the vowel. To this point, the proposed model for speech coarticulation has been, consistently, a time-variant model without dependence upon a framework of discrete acoustic units.

Indeed, an acoustic waveform segmentation between a vowel and its adjacent *stop* consonant has precedence in the literature and can be performed with some reliability. However, the two other classes of consonants examined in this study, nasals and liquids, cannot be segmented as easily from the adjacent vowel. Nasals and liquids are categorized in a feature class known as sonorants (Ladefoged 1975, p. 239). The sonorant prime feature class includes all vowels. It specifies those phonemes which yield a high level of acoustic energy. Nasals and liquids can extend over relatively long durations, like vowels. And, because of their high acoustic energy level, a nasal or liquid can be well-integrated with the adjacent vowel. A precise segmentation between a vowel and sonorant consonant is likely to be *un*reliable.

Therefore, for the purposes of the present study, the signal $z2(t)$, which is to be associated with the contextual vowel C/V/C, was recorded from the *entire* CVC utterance. No acoustic waveform editing of the CVC signal occurs. Instead, the task of discriminating the vowel from the adjacent consonants is delegated to the **COAR**(a,b) function. This is possible because the time-dependent attribute of the **COAR**(a,b) allows it to discriminate between various time-localized events.

Consider, then, at the initial and final margins of time b, the **COAR**(a,b) distribution will contain cross-wavelet correlations formulated between the isolated vowel and each *consonantal* portion of the CVC. In other words, at each end of the CVC, the cross-wavelet attempts to correlate /V/ with /C/.

Furthermore, in those cases where a high degree of CVC coarticulation may be present, it is expected that a *gradual* change in the **COAR**(a,b) distribution will be manifested over b. The gradual change occurs by virtue of the continuous transition

from initial /C/ to the coarticulated /V/ to final /C/. On the other hand, in those cases of testing **COAR**(a,b) using the stops (/b/, /d/, and /g/), it is expected that a very low magnitude[**COAR**(a,b)] will result at the endpoints associated with either stop consonant. This is because, at those time-localized portions of the **COAR**(a,b), a correlation is posed between the isolated vowel /V/ and a stop consonant /C/. Such a combination should not constitute, in theory, a favorable correlation.

Chapter 7

RESULTS

This chapter presents the results obtained from the experimental study. Here, the results refer to the body of calculations derived from samples of recorded speech. Each calculated data object appears in the form of a two-dimensional matrix of transform coefficients.

The transform coefficients were calculated from either the wavelet transform integral, $\mathbf{W}_f z$ (equation [3.1]), or the *cross* wavelet channel estimate, **[CÔAR]** (equation [5.7]). More specifically, a transform matrix contains the magnitude of the integral coefficient taken as a function of a time parameter (b) and a scale/frequency parameter (a). Although it has previously been expressed as a function of (s,τ), the **[CÔAR]** function appears in this chapter as **[CÔAR]**(a,b).

The transform matrices are illustrated in the form of three-dimensional plots, whereby, the horizontal axis depicts variation over time, and the vertical axis depicts variation in scale/frequency. A continuous gray-scale in the plot depicts the magnitude of the coefficient at each location in time and scale. Black areas in the plot indicate large-magnitude values; gray areas represent intermediate values. The white portions of the plot indicate values of the matrix which fall below the small-magnitude threshold or "noise floor".

Five types of plots are presented in this chapter. The first type is the Morlet wavelet transform of a single utterance. The Morlet wavelet transform illustrates the

distribution of spectral energy for that utterance as a function time and frequency. A series of such plots is followed by another series of *cross* wavelet plots. The cross wavelet plot is the **[CÔAR]**(a,b) distribution estimate (a particular instance of evaluation for the coarticulation model). The **[CÔAR]**(a,b) plot is derived from a [/V/, C/V/C] utterance pair, and it illustrates the distribution which plays the role of the "coarticulation channel" estimate.

The third type of plot appearing in this chapter is a *modified* version of the (previous) wavelet transform. This is followed by a modified version of the **[CÔAR]**(a,b) distribution, the "windowed **COAR**(a,b)". It will be shown that the modified versions of these constructs are necessary variations from the original, and that they yield a superior representation of the coarticulation channel.

The final type of plot to be presented here is the classical spectrogram of a lone utterance. A series of spectrograms are shown alongside their counterpart **[CÔAR]**(a,b) distributions. This combination serves as a means of direct comparison between the classical illustration of CVC coarticulation and the proposed coarticulation model.

7.1 Wavelet Transform Results

The following pages illustrate wavelet transform distributions calculated for some example utterances. Each of the utterances were spoken in isolation. (The final-position stop consonants were consistently exploded.) The mother wavelet is the Morlet. The plot shows the magnitude of the wavelet coefficient as a function of log frequency (kHz) and time (milliseconds).

The darkest areas in the plot depict regions of high magnitude (0 dB). The white areas extend down in magnitude to −40 dB. The time axis (horizontal) is evaluated every 2.496 milliseconds. Evaluations in frequency (the vertical axis) are spaced geometrically (log spacing), and these occur as a regular factor of 1.037. The number of evaluations is 400 in time and 100 in frequency (scale).

Figure 7.1 plots the wavelet transform of the isolated vowel /u/. The plot depicts a series of horizontal bars which correspond to the harmonics of the voiced excitation. (These bars would be evenly spaced on a linear frequency scale.) It should be noted that the lowest of these horizontal bars is *not* the fundamental voicing frequency. The subject's fundamental frequency is approximately 90 Hz; however, these wavelet transforms begin at 110 Hz. The lowest group of horizontal bars (approximately four) is spanned by the *F1* vowel formant which is centered about 300 Hz. The next highest resonance, *F2*, appears at about 900 Hz (just below log frequency = 0). The horizontal band second from the top indicates *F3*, which resides at a frequency about 2250 Hz.

The wavelet transform plot for the CVC utterance /dud/, shown in Figure 7.2, maintains the same basic formant structure as the isolated /u/ utterance. However, two sharp vertical stripes, indicating the stop-burst transients at /d/ initial and final, can be identified at times −175 ms and +250 ms, respectively. A voicing gap which immediately precedes the final /d/ stop-burst is also visible. Finally, a dynamic parabolic trajectory in the *F2* formant can be identified. This vowel formant is exhibiting a time-varying coarticulation effect, attributable to both initial and final /d/ consonants.

Also apparent from this wavelet transform plot is the variable time-frequency resolution of the Morlet wavelet transform. Notice that individual harmonics can be

A Morlet Wavelet Transform of /u/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

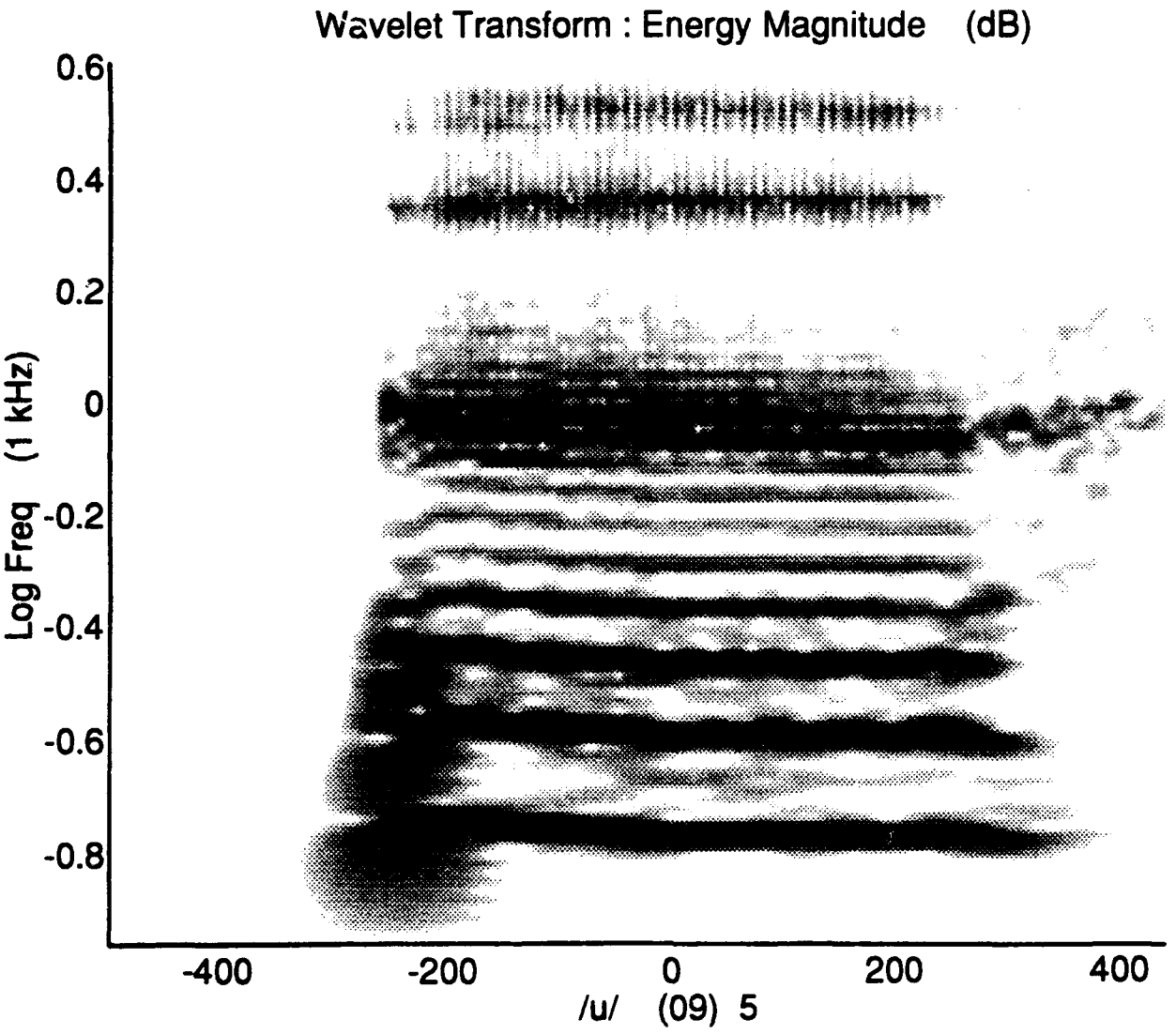


Figure 7.1 A Morlet Wavelet Transform of /u/

A Morlet Wavelet Transform of /dud/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

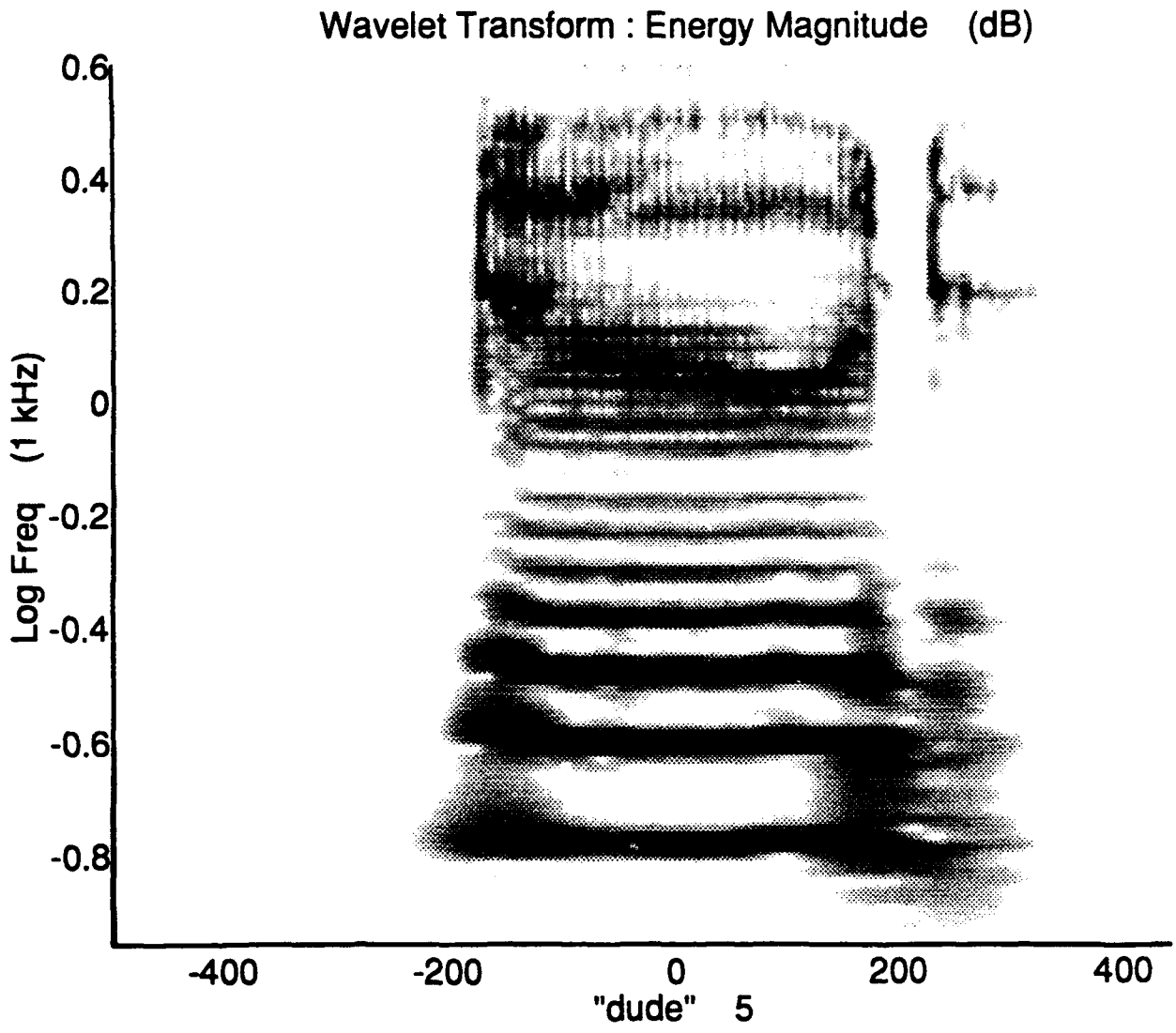


Figure 7.2 A Morlet Wavelet Transform of /dud/

easily discerned at low frequencies, whereas, the same harmonic structure is blurred at higher frequencies. This reflects the superior frequency resolution of the wavelet transform at low frequencies. On the other hand, consider the events of /dud/ in the time-domain, such as the initial and final bursts and the voicing gap. These events are sharply defined at higher frequencies, yet, they become blurred at lower frequencies. This trade-off is consistent with the wavelet transform's superior time resolution at high frequencies.

The third wavelet transform plot shown in Figure 7.3, /rär/, indicates a basic vowel formant structure appropriate for the vowel /ä/: $F1 \approx 660$ Hz, $F2 \approx 1020$ Hz, and $F3 \approx 2240$ Hz. A weak burst of voicing onset is apparent in the mid-frequency vertical stripe located about time -200 ms. A noteworthy feature of this plot, however, appears in the dynamic formant structure. Notice the concave downward parabolic trajectory on *F1*, the concave upward trajectory on *F2*, and the concave downward trajectory of *F3*. These formant trajectories indicate a fluid coarticulation from /r/ to /ä/ to /r/, whereby, the medial "target" formant values for /ä/ (occurring at about the time 0 ms) are sustained over only a minor portion of the vowel's duration.

As an illustration of how this wavelet transform representation differs from the classical spectrogram, consider the remaining three figures. Figures 7.4, 7.5, and 7.6 are sets of wavelet transforms calculated from a selection of utterances from the word list.

Figure 7.4 gives a cross-sectional view for four different C/−/C contexts around the same vowel /u/: /gug/, /nun/, /rur/, and /lul/. Notice that the high frequency portion of each wavelet representation resembles a wideband (300 Hz) spectrogram,

A Morlet Wavelet Transform of /rär/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

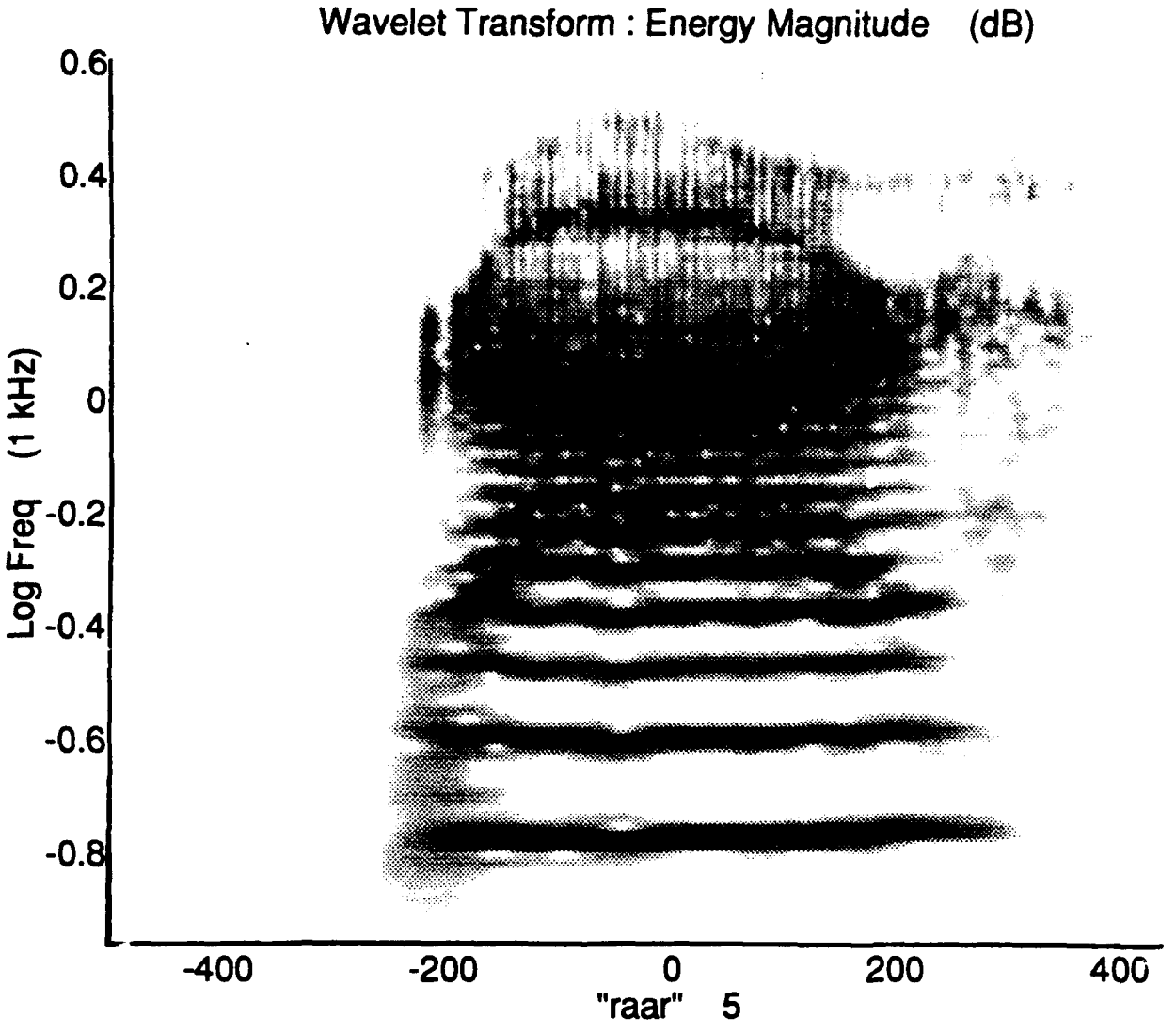


Figure 7.3 A Morlet Wavelet Transform of /rär/

Wavelet Transforms of some /u/ words: /gug/, /rur/, /lul/, /nun/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

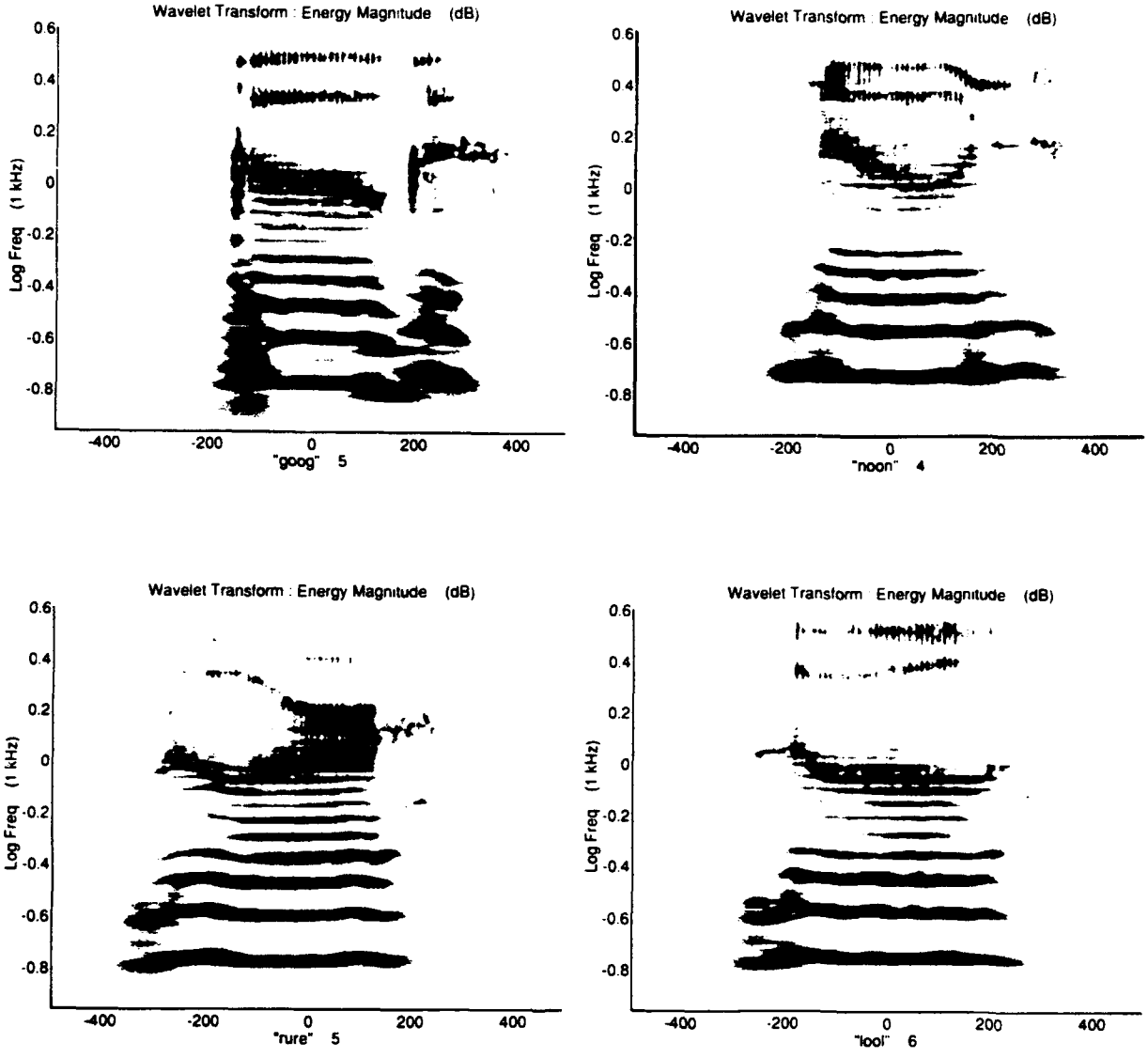


Figure 7.4 Wavelet Transforms of some /u/ words: /gug/, /rur/, /lul/, /nun/

having good time resolution with a pattern of vertical striations. Conversely, the low frequency portion of the wavelet representation resembles a narrowband (45 Hz) spectrogram, having good time resolution and a pattern of horizontal striations. Attributes from both types of spectrograms, therefore, appear within a single wavelet representation. Naturally, the advantages and disadvantages of each are maintained within their respective "half-planes".

Notice also from Figure 7.4 that the minimum frequency boundaries of these c/u/c utterances are well defined. In other words, the lowest order harmonics of the vowel are highly resolved. This allows the lower boundary of the *F1* formant to be distinguished from the fundamental frequency. Within these wavelet representations, the frequency region containing the fundamental has been omitted; however, the very *first* harmonic ridge is visible and fully resolved. In the case of a spectrogram, good separation between *F1* and the fundamental is not always achieved, particularly for a "high" vowel such as /u/ (for which *F1* reaches a minimum value). (See Figure 7.17.)

Figure 7.5 shows the four vowels, /i/, /æ/, /ä/, and /u/, in the context of the bilabial stop consonant: b/−/b. Consider how the stop bursts in these utterances have been resolved in time by the wavelet representation. As previously stated, the time-resolution of an impulsive burst varies with frequency. However, it can be assumed that an impulsive articulatory event, such as the release in initial /b/, or the closure in final /b/, is an event which is *synchronous* in time. In other words, the energy onset/offset occurs *simultaneously* for all of the available frequencies. Therefore, the superior time resolution at high frequency may be "extrapolated" down to the lower frequency regions.

Wavelet Transforms of the /b/ words: /bib/, /bæb/, /bäb/, /bub/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

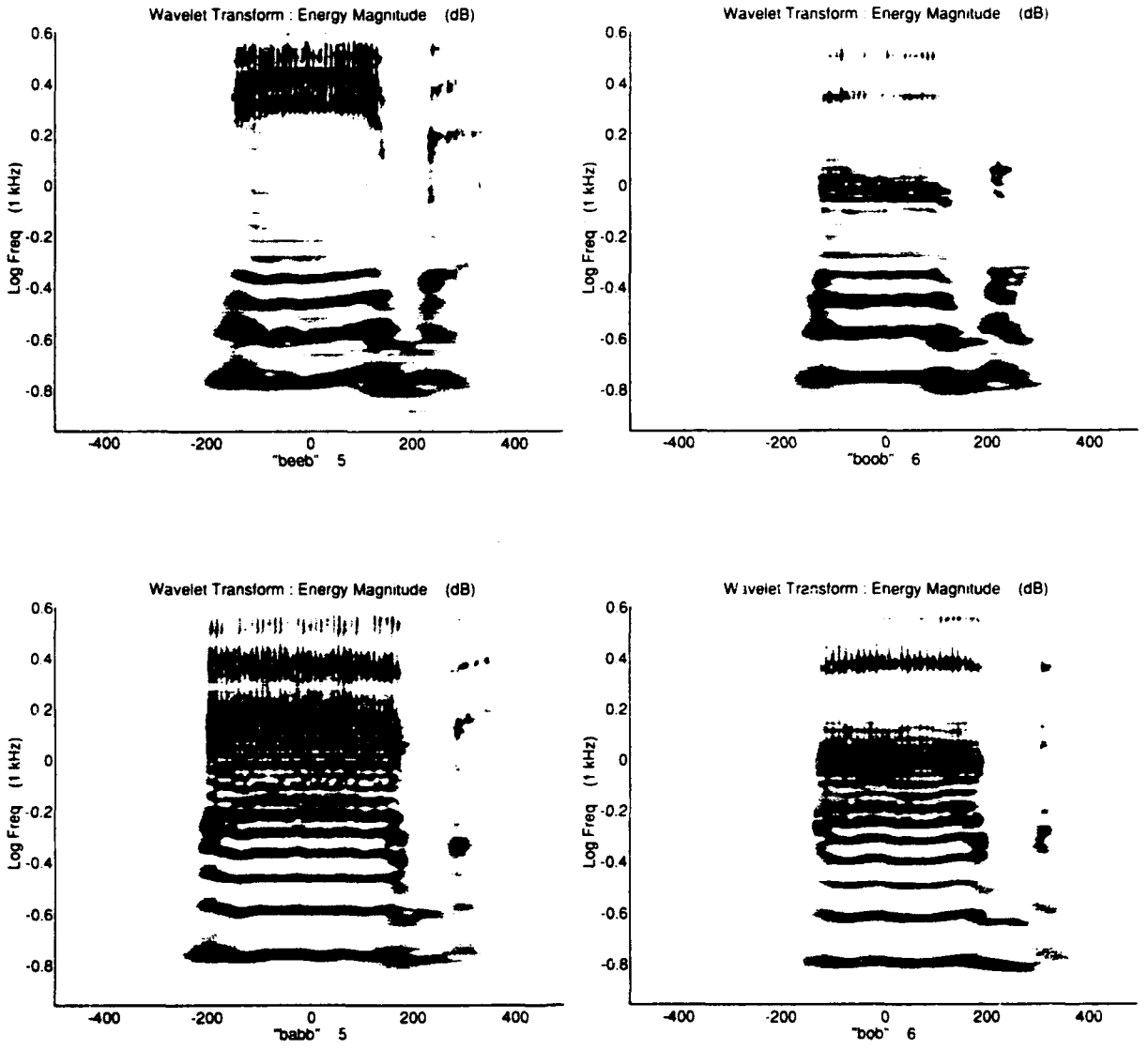


Figure 7.5 Wavelet Transforms of the /b/ words: /bib/, /bæb/, /bäb/, /bub/

Using this method, certain articulatory events can be very precisely pinpointed in time. Observe, for example, in Figure 7.5:

1) The final burst in "beeb" (at time +250 ms)

2) The initial burst in "babb" (at time −200 ms)

3) The voicing-gap (voice onset time) in "bob" (+200 to +325 ms).

As a final illustration of how these wavelet representations differ from the classical spectrogram, consider nasalization. Figure 7.6 shows the four vowels in the company of the bilabial nasal stop: m/−/m. These plots convey a large quantity of information about these utterances in a balanced and detailed manner. In particular, the wavelet transform provides a favorable contrast between each vowel and the final /m/. Given that each of the three phones are sustained over a substantial time duration, and that each has a complex formant/spectral structure, their contrasting differences are manifested at numerous locations throughout the time-frequency plane. (As an example of how one of these utterances, "moom," would appear on a spectrogram, refer ahead to Figure 7.20.)

For example, in the case of the word "mom," the difference between the vowel and final /m/ is shown as a sudden drop in the overall intensity level. The word "meem" exhibits a similar energy contrast over the low-frequency half-plane. In the upper half-plane of "meem," however, a *displacement* in the formant structure is apparent. The upper and lower half-planes in "moom" can also be divided. In this case, the upper frequency region maintains the structure and increases the energy of formants (from /u/ to /m/); whereas, in the lower frequency region of "moom," the familiar *F1* attenuation appears.

Wavelet Transforms of the /m/ words: /mim/, /mæm'/, /mäm/, /mum/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

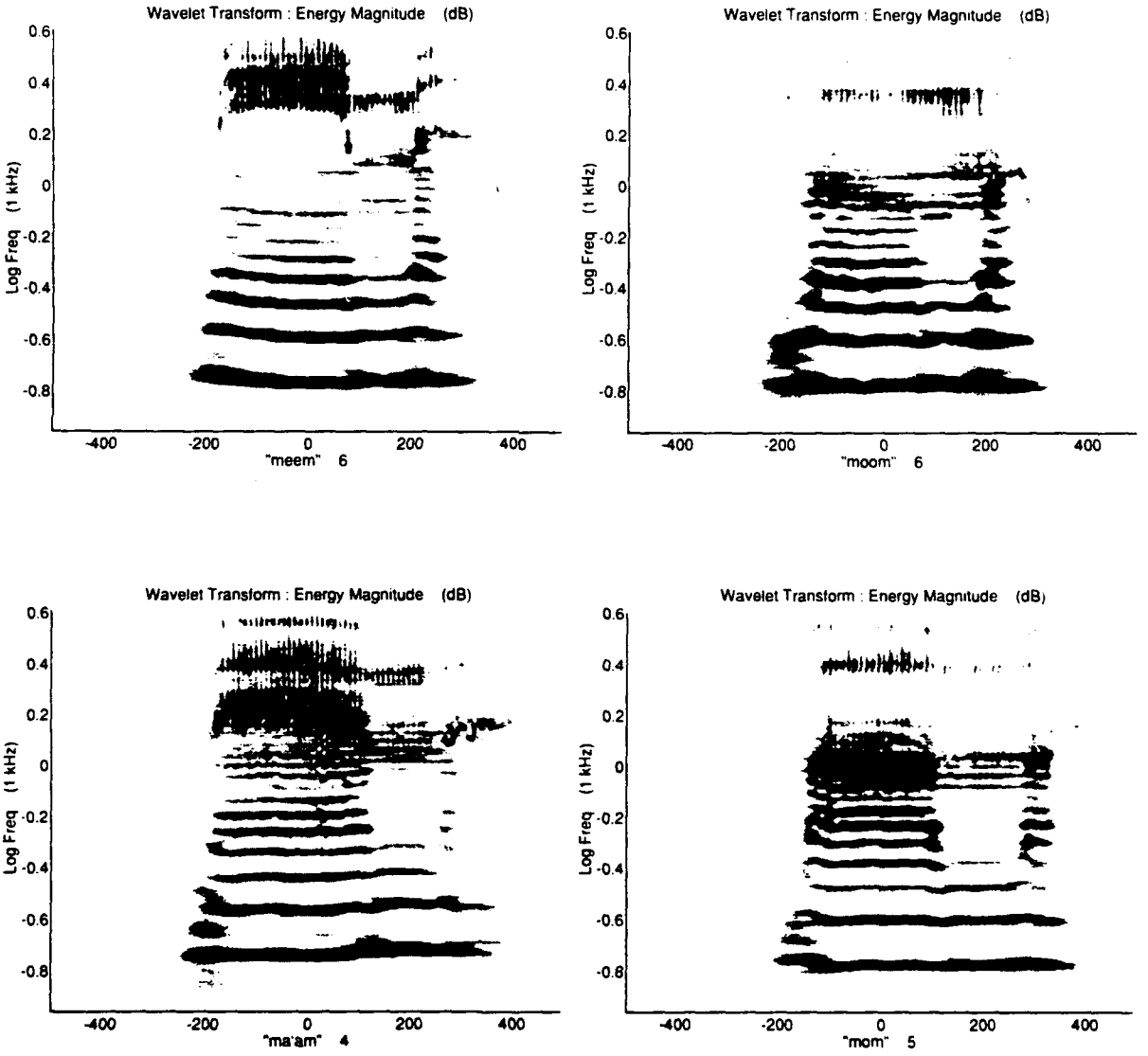


Figure 7.6 Wavelet Transforms of the /m/ words: /mim/, /mæm/, /mäm/, /mum/

These wavelet transform plots are favorable representations of the m/V/m words in the sense that all portions of the transform plane are utilized to convey the significant attributes of these utterances. Conversely, the m/V/m utterance exhibits variations in the time, frequency, and intensity domains, which are compatible with the full range and resolution of these wavelet transforms.

7.2 Cross Wavelet Transform Results

This section presents the **[CÔAR]**(a,b) channel estimates calculated for a series of [/V/, c/V/c] utterance pairs. Included in the following group of plots are the coarticulation channel estimates for the vowels [/i/, /æ/, /ä/, /u/] into four different consonantal contexts [b/−/b, d/−/d, m/−/m, and r/−/r]. Figure 7.7 shows how the cross wavelet calculations appearing in these plots relate to the theoretical model presented previously (see Figure 4.2).

The magnitude of the **[CÔAR]**(a,b) coefficient is plotted as a function of −log[scale] and time-shift (milliseconds). Figure 7.8 shows a **[CÔAR]**(a,b) channel estimate for the utterance pair [/u/, d/u/d].

The darkest areas in the plot depict regions of high magnitude (0 dB). The white areas extend down in magnitude to −40 dB. As in the case of the wavelet transform shown previously, the time axis (horizontal) is evaluated at 2.496 millisecond intervals. Evaluations in scale (the vertical axis) have a logarithmic spacing, at a regular factor of 1.037. There are 800 evaluations in time (b) and 40 evaluations in scale (a).

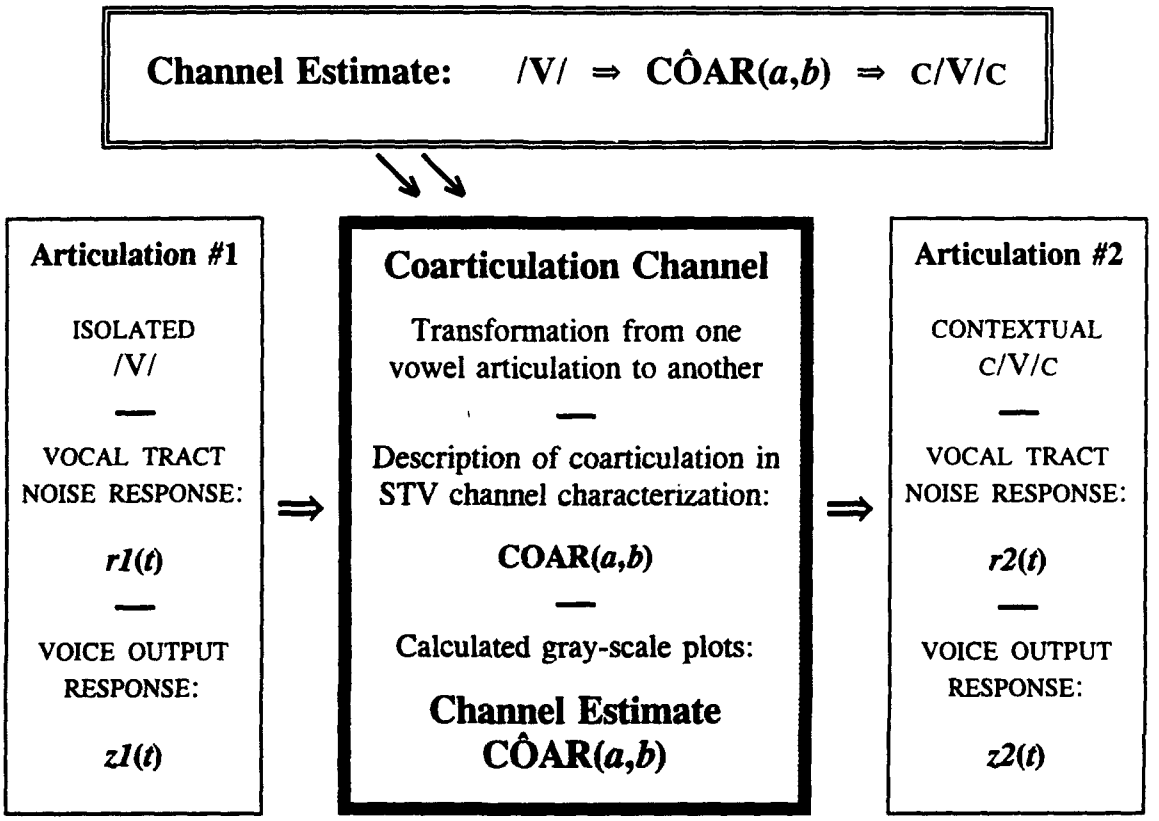


Figure 7.7 **The Calculated Channel Estimate [CÔAR](*a*,*b*)**

Channel Estimate: /u/ ⇒ CÔAR(a,b) ⇒ d/u/d

"dude"

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

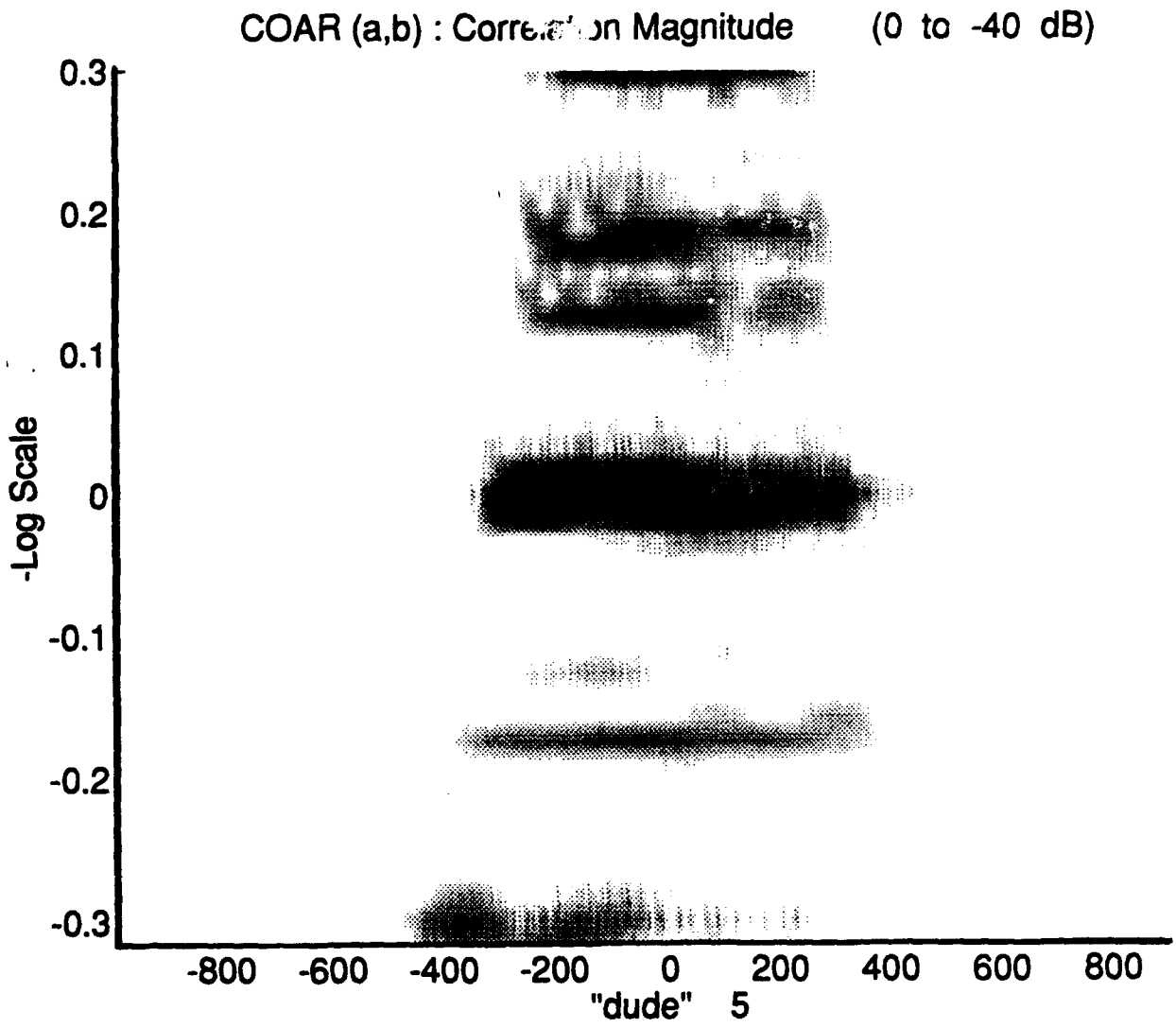


Figure 7.8 Channel Estimate: /u/ ⇒ CÔAR(a,b) ⇒ d/u/d

However, for the purposes of plotting within the figure, every other time point has been averaged with its adjacent point. This yields a reduced *graphic* representation of 400 time-points by 40 scale-points.

The time-shift range extends from −1000 ms (left) to +1000 ms (right). The scale range extends from 0.5 (top) to 2.0 (bottom). The orientation of the scale axis is such that the compressed perturbations (*more* zero crossings) appear at the top, whereas, the dilated perturbations (*less* zero crossings) appear below.

Figure 7.9 shows the **[CÔAR]**(*a*,*b*) distributions for each of the four vowels taken in their b/−/b context. The figure contains cross wavelet transforms for [/i/, "beeb"], [/æ/, "babb"], [/ä/, "bob"], and [/u/, "boob"]. Each plot depicts a correlation of the various (scaled and shifted) versions of /V/ with the (unperturbed) CVC. The orientation is such that:

1) At the *left* time region of each **[CÔAR]**(*a*,*b*) plot, /V/ is time-shifted relative to the CVC, whereby, /V/ comes *before* the CVC.

2) At the *right* side of the time-shift axis, /V/ is time-shifted in such a way that it occurs *later* than the CVC.

3) At the *top* scale region of the **[CÔAR]**(*a*,*b*) plot, /V/ is *compressed* in scale relative to the CVC.

4) At the *bottom* portion of the scale axis, /V/ is *dilated* in scale relative to the CVC.

Refer to the Figure 3.2 shown previously. Consider the following operations of scaling and shifting a vowel according to the terms posed by that illustration:

Channel Estimate: /V/ ⇒ $\hat{COAR}(a,b)$ ⇒ b/V/b

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

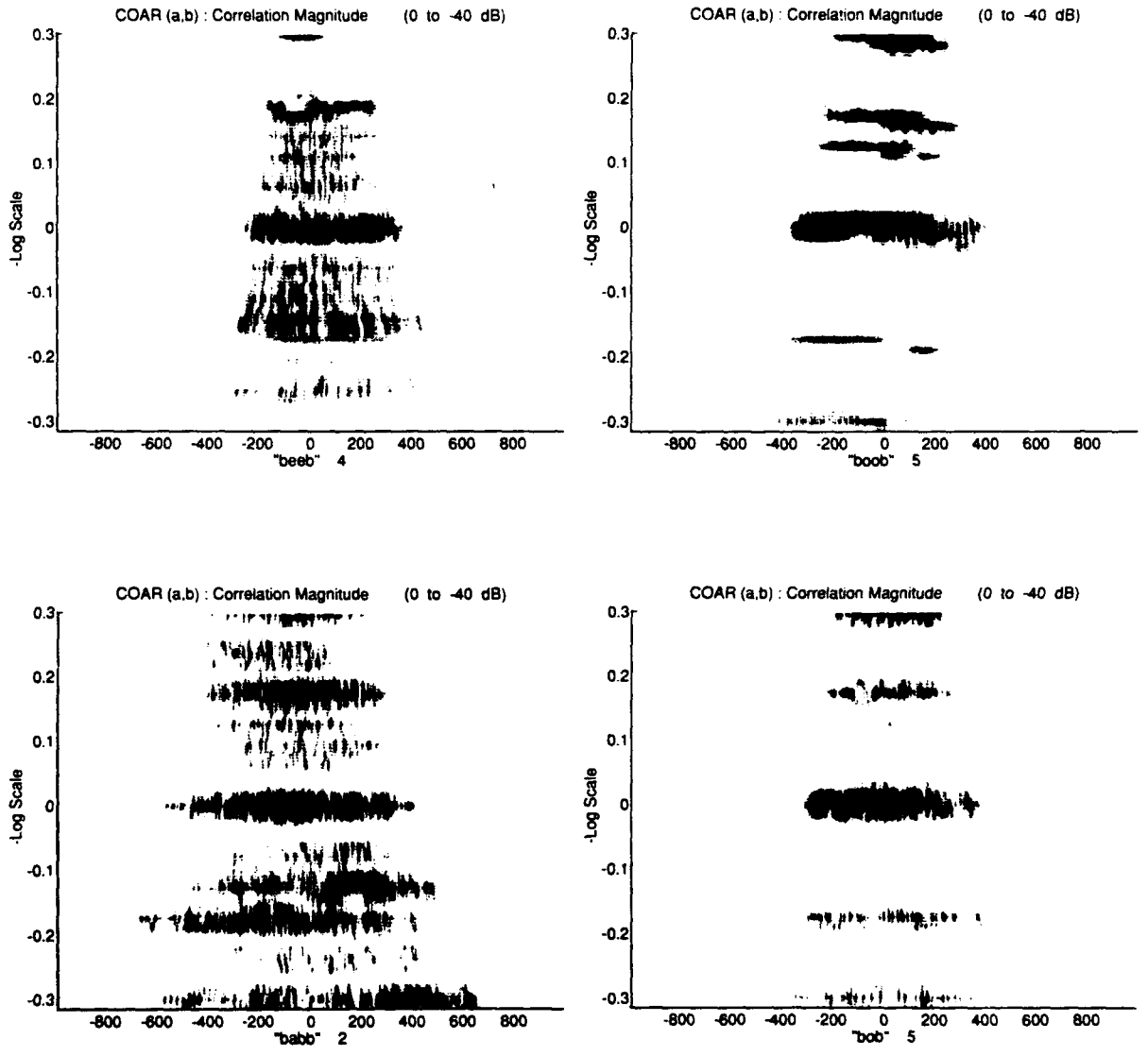


Figure 7.9 Channel Estimate: /V/ ⇒ $\hat{COAR}(a,b)$ ⇒ b/V/b

1) The /V/ that is perturbed in the manner of $(a,b)=(4,-17)$ maps into the lower-left region of the **[CÔAR]**(a,b) plot. The intensity of gray-scale there reflects the correlation of an *early*, *dilated* /V/ with the CVC.

2) The /V/ that is *un*perturbed, $(a,b)=(1,0)$, maps into the central region of the **[CÔAR]**(a,b) plot. The intensity of gray-scale there reflects the correlation of a *middle*, *normal* /V/ with the CVC.

3) Finally, the isolated vowel perturbed in the manner of $(a,b)=(0.25,6)$ maps into the upper-right region of the **[CÔAR]**(a,b) plot. The intensity of gray-scale there reflects the correlation of a *late*, *compressed* /V/ with the CVC.

Notice that in all of the plots of Figure 7.9, the highest concentration of energy occurs in the central region, ranging roughly from -300 to $+300$ ms in time; -0.02 to $+0.02$ in $-$log scale. This means that the *un*scaled ($a = 1.0$) version of the isolated vowel is most similar to the CVC version. It also means that when the /V/ is shifted in time along the CVC, a high correlation results whenever the two signals overlap in time.

The regions of these plots *away* from the origin contain some grayish areas of mid-level energy. This means that many various versions of the perturbed /V/ exhibit moderate levels of similarity with the CVC. On the other hand, in those regions where the **[CÔAR]**(a,b) distribution shows pure white, the associated time-shift yields a vowel which has been shifted too early or too late. In these regions, no overlap occurs between the signals associated with the isolated vowel and the CVC.

7.3 The Role of the Vowel's Self Similarity

In examining the individual **[CÔAR]**(a,b) distributions shown in Figure 7.9, it appears that the [/i/, "beeb"] cross wavelet (shown as "beeb") exhibits a rather diffuse correlation. This is indicated by the continuous pattern of gray striations distributed over the middle portion of the (a,b) plane. In contrast, the [/ä/, "bob"] cross wavelet (shown as "bob") is characterized by a series of well-defined, stark horizontal stripes. The same is true for the [/u/, "boob"] pair (shown as "boob").

For these vowels showing stark contrasts with respect to various scale values, the **[CÔAR]**(a,b) function exhibits some *selectivity* in scale. In the [/ä/, "bob"] plot, for instance, the isolated /ä/ is highly similar to the −/ä/− within "bob," at certain *specific* scale values (the dark ridges). At other scale values (the stark white patches), the isolated /ä/ is decidedly *dis*similar to the −/ä/− within "bob". Because these patterns are dominant over the time interval on which /ä/ and −/ä/− are directly aligned, this selectivity in scale is associated with the similarity of the vowel /ä/ to *itself*. In other words, /ä/ is highly similar to itself when scaled, but at a very particular set of scale values. Referring to the plot in Figure 7.9, the peaks of "self-similarity" occur at the following −log scale values: −0.3, −0.19, 0.0, +0.19, and +0.3. (A maximum degree of self-similarity is expected at −log scale = 0, because there the vowel is neither dilated nor compressed.)

The same applies for the vowel /u/ in the [/u/, "boob"] plot. The set of high-contrast horizontal bars appearing in that plot indicate a distinct pattern of self-similarity for /u/. In signal processing terms, the cross wavelet transform between any signal and

itself (such as that associated with a spoken vowel) is known as the "auto-ambiguity" function.

One possible explanation for the patterns of self-similarity observed in these utterances is an interaction between the vowel's own formant frequencies. For example, the cross wavelet transform between two repetitions of the vowel /ä/ results in a coupling between *F1* of one repetition with *F2* of the other. Consider the *F1* component of /ä/ when compressed by the affine mapping. At some value of scale < 1.0, the compressed *F1* will correlate favorably with the *F2* component. Indeed, the *ratio* between the *F1* and *F2* frequencies for the vowel /ä/ is roughly (660 Hz/1020 Hz) = 0.65. The log(0.65) = 0.19 −log scale, and this is the approximate location of the upper horizontal stripe appearing in the [/ä/, "bob"] cross (Figure 7.9). The lower stripe, located at −0.19 −log scale, may be associated with the interaction between *F2* of one repetition and *F1* of the other. I.e., the *inverse* ratio is: $F2/F1 \approx 1.55 \approx \log^{-1}(0.19)$.

This explanation for the observed patterns of vowel self-similarity (that which relies on vowel formant interaction) is not always appropriate, however. Consider another example, the [/u/, "boob"] cross of the same Figure 7.9. Here, two distinct horizontal ridges are visible on the upper portion of the plot. They are located at −log scale 0.13 and 0.18. (The lower horizontal ridge, at −0.18 −log scale, is most likely the inverted or "mirror" version of the upper peak). None of these ridge locations, however, can be explained by the interaction between /u/ formants. All three peaks occur at scale values somewhere between 0.5 and 2.0 (the limits in scale for all of the cross wavelet plots). Yet, the /u/ formant frequencies (300, 900, and 2250 Hz) are well-separated in frequency. The nearest two neighbors among these formants generate

frequency ratios of 0.4 and 2.5. *If* an interaction between the formants of /u/ were to be manifested in its auto-ambiguity (the cross wavelet of /u/ with /u/), then the resultant ridge peaks would appear at scale values *outside* of the range administered in these calculations.

In summary, it is not likely that the patterns of vowel self-similarity observed in these plots can be attributed to a simple interaction, namely, the mutual reinforcement of the vowel's formant peaks. What can be shown, however, is that such patterns are indeed characteristic aspects of the *vowel*. This will be evident from the following set of cross wavelet plots.

Figure 7.10 shows the **[CÔAR]**(a,b) distributions for each of the four vowels taken in their m/−/m context. The figure contains cross wavelet transforms for [/i/, "meem"], [/æ/, "ma'am"], [/ä/, "mom"], and [/u/, "moom"]. As before, the highest concentration of energy occurs in the central region, in a horizontal band centered about unity scale. Notice that the highly diffuse patterns of vowel self-similarity in /i/ and /æ/ (shown as "meem" and "ma'am," respectively) have been repeated from Figure 7.9. These plots stand in contrast to [/ä/, "mom"] and [/u/, "moom"] which are relatively more compact and contain well-defined peaks.

Notice also from Figure 7.10 that the locations of the horizontal ridges have *not* shifted from their respective positions in Figure 7.9. In other words, the most prominent ridge peaks appear at the same scale values in the m/−/m context as they did in the b/−/b context. This is true for each of the four vowels, including the diffusely patterned /i/ and /æ/ vowels.

Channel Estimate: /V/ ⇒ CÔAR(a,b) ⇒ m/V/m

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

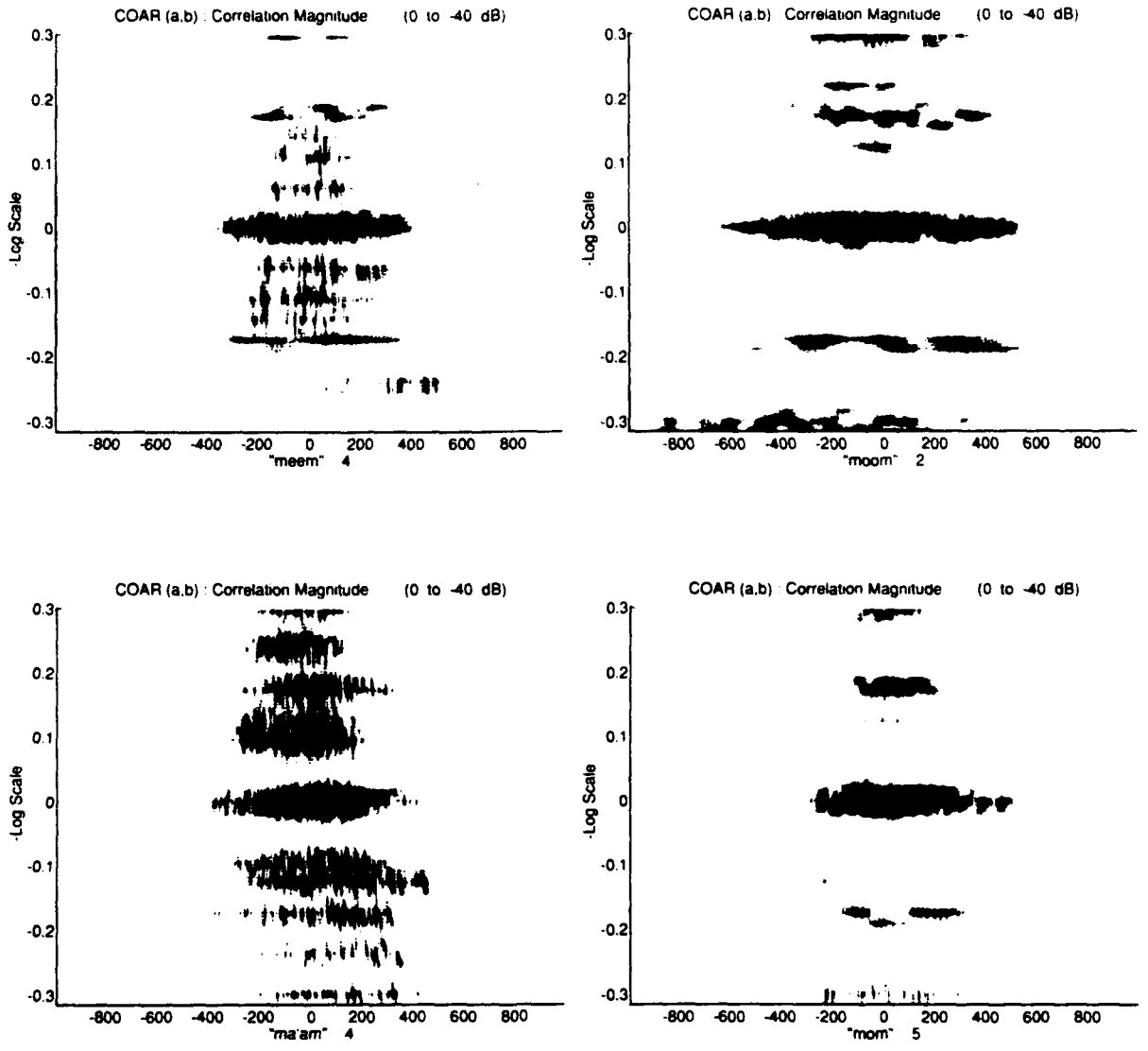


Figure 7.10 Channel Estimate: /V/ ⇒ CÔAR(a,b) ⇒ m/V/m

The consonantal context in c/−/c, therefore, does not alter the pattern of self-similarity observed for any of the vowels. Rather, the same vowel patterns manifested in the [/V/, b/−/b] crosses are maintained for other c/−/c contexts, with good reproducibility. (The plots of Figure 7.11 may also be consulted in this regard.)

7.4 The Lack of Time Variability in the COAR Distribution

Consider any of the **[CÔAR]**(a,b) estimates of the previous plots and notice their limited variability in time (b). For only a few of the plots does there appear to be much time-variation. For example, the distributions "boob" (Figure 7.9), "mom," and "moom" (Figure 7.10), undergo slight changes in direction or intensity level as a function of the time-shift parameter. For the most part, however, the **[CÔAR]**(a,b) estimates, which are designed to model the time-dynamic processes of coarticulation, are *static* functions of time.

This uniformity with respect to time can be further observed in the plots which follow. Figure 7.11 shows the **[CÔAR]**(a,b) distributions for the vowels taken in their r/−/r context. The figure contains the cross wavelets for [/i/, "rear"], [/æ/, "ræ r"], [/ä/, "raar"], and [/u/, "rure"].

It is expected that a dynamic model of CVC coarticulation should be capable of depicting some of the time-dependent transitions which are undoubtedly contained in these CVC signals. There is good reason, however, for the model's limitation in this respect.

Channel Estimate: /V/ ⇒ CÔAR(a,b) ⇒ r/V/r

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

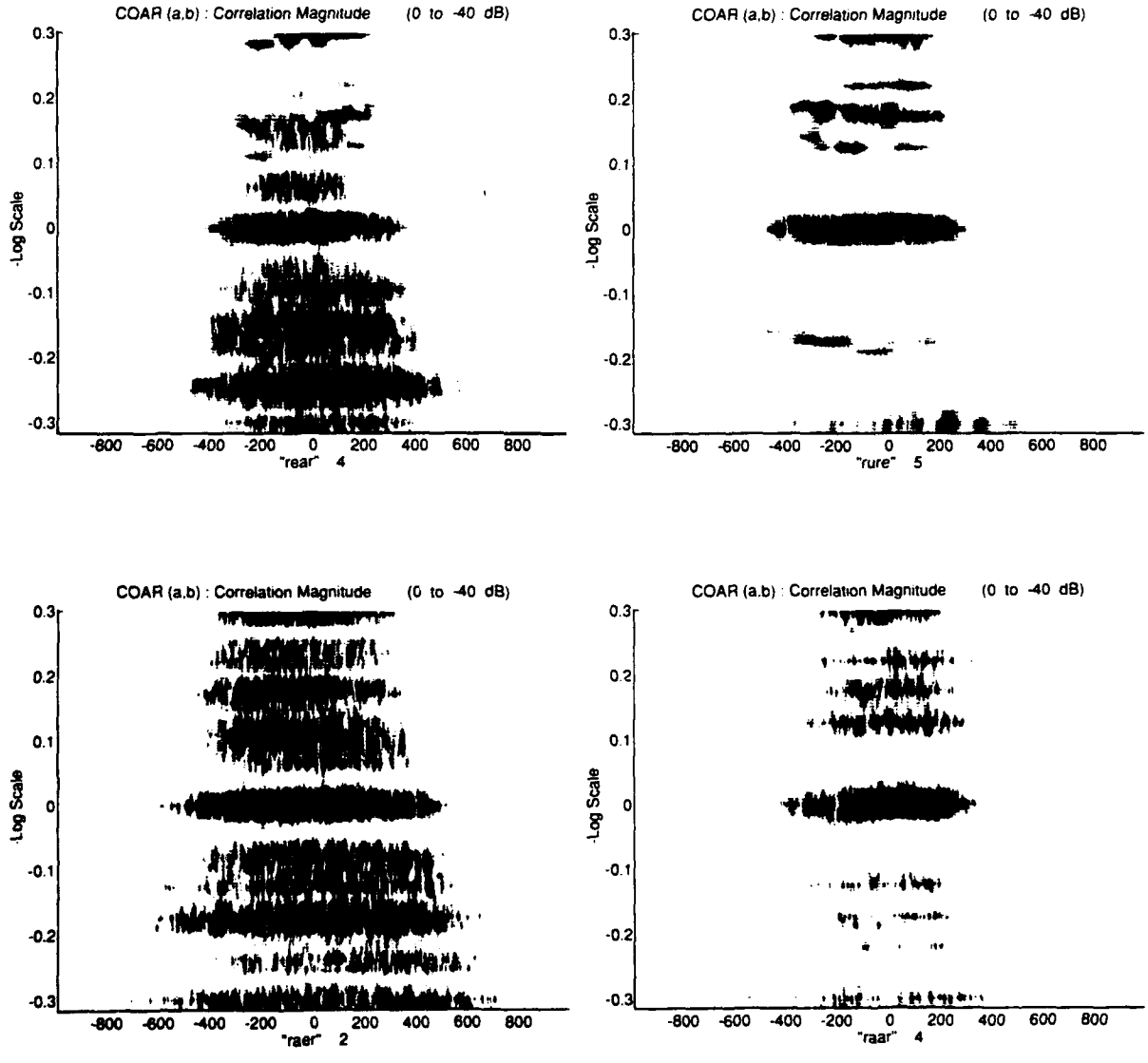


Figure 7.11 Channel Estimate: /V/ ⇒ CÔAR(a,b) ⇒ r/V/r

Consider that the signal recorded from the isolated vowel is sustained over the same (long) time duration as that recorded from the CVC utterance. In other words, the isolated vowel and the CVC have roughly the same time length. The coarticulation channel estimate, **[CÔAR]**(a,b), is calculated as the wavelet transform of the CVC signal, using the isolated vowel as the mother wavelet. In signal terms, therefore, the analysis wavelet has the same "time-support" as the signal being analyzed. Since the wavelet used to analyze the signal occupies the same stretch of time as the signal itself, the wavelet transform's ability to resolve time-varying features in the signal is critically limited.

Instead of portraying the signal's dynamic/transient behavior, this type of wavelet transform is apt to identify the signal's *general location* in time; it shows "dark" whenever the signals overlap and "white" whenever they do not. When the time-shift parameter brings the (long) wavelet into alignment with the signal, the resulting distribution is, in effect, a "time-averaged" measure of the signals components at different scales. The long analysis time-window effectively treats the entire duration of the signal as a single event.

Many of the **[CÔAR]**(a,b) distributions presented thus far appear to fit this description. Given this lack of time variability in the **COAR**(a,b) model, the remedy is to *reduce* the time-support of the analysis wavelet (the signal associated with isolated /V/). This is achieved by applying a time window to the wavelet transform of the isolated /V/.

7.5 Time Windowing the Wavelet Transform of the Isolated Vowel

The **COAR**(a,b) cross-wavelet distribution has good time resolution if the analyzing part of that correlation is well-*localized* in time. In other words, the time-support of the isolated vowel must be small relative to that of the CVC utterance. In particular, the isolated vowel must have a duration on the order of the CVC's fastest transitions. If the isolated vowel is restricted to an interval of 10 or 20 milliseconds, in such a way that its spectral structure is maintained, then the vowel can be well-suited for tracking consonant/vowel transitions over the course of the CVC. When calculated using a "short" isolated vowel representation, therefore, the **COAR**(a,b) reflects (in b) the dynamic characteristics of the CVC coarticulation.

Recall that the **[CÔAR]** estimate is obtained in this study by performing a mother mapper operation on two other wavelet transforms:

$$[5.7] \qquad \mathbf{[C\hat{O}AR]}\,(s,\tau) \;=\; \frac{1}{C_x}\int \frac{1}{a^2}\int \mathbf{W}_{f(t)}\,z2 \cdot \mathbf{W}^{*}_{f(t)}\,z1\left(\frac{a}{s},\,\frac{b-\tau}{s}\right) db\; da$$

where $z1(t)$ is the recorded microphone signal of the isolated /V/ utterance, and $z2(t)$ the recorded signal of the CVC utterance. This method of **[CÔAR]** estimation calculates the wavelet transform of the isolated vowel $[\mathbf{W}_f z1(a,b)]$, and it calculates that of the contextual vowel $[\mathbf{W}_f z2(a,b)]$. The coefficient matrix associated with the wavelet transform of the isolated vowel is therefore available at this intermediate stage. This matrix provides an opportunity for *windowing* the vowel in such a way that its spectral structure is maintained.

A smooth window is applied in the time-shift (b) domain *to* the wavelet transform coefficients [$W_{fzl}(a,b)$]. (The window function is constant with respect to scale.) The time-support of the isolated vowel representation is effectively reduced. Notice, however, that by windowing the wavelet *transform* coefficients, the spectral distortions normally associated with *signal* windowing (such as spectral leakage and scalloping losses) are nicely avoided (DeFatta et al. 1988, section 6.6).

Consider, furthermore, that an isolated vowel is (primarily) a sustained, steady-state articulation. During the medial portion of such a vowel, therefore, the wavelet transform distribution $W_{fzl}(a,b)$ is relatively *constant* over time. The time-window then aligns to capture a small, *static* interval over the medial portion of the vowel. The resulting representation thus contains most or all of the relevant spectral information for that vowel.

Figure 7.12 shows the wavelet transforms calculated from a selected set of isolated vowels: /i/, /æ/, /ä/, and /u/. Figure 7.13 shows the time-windowed versions of these *same* wavelet transforms. The distributions plotted in the latter figure utilize, as a start, the exact distributions appearing in Figure 7.12. The applied window is a Gaussian weighting function in b. The 3 dB time-width of these windows is 20 ms. (I.e., the 3 dB attenuation point is 10 ms away from the 0 dB midpoint). Notice that the Gaussian window is consistently centered at some medial point in the vowel.

Wavelet Transforms of the 4 isolated vowels: /i/, /æ/, /ä/, /u/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

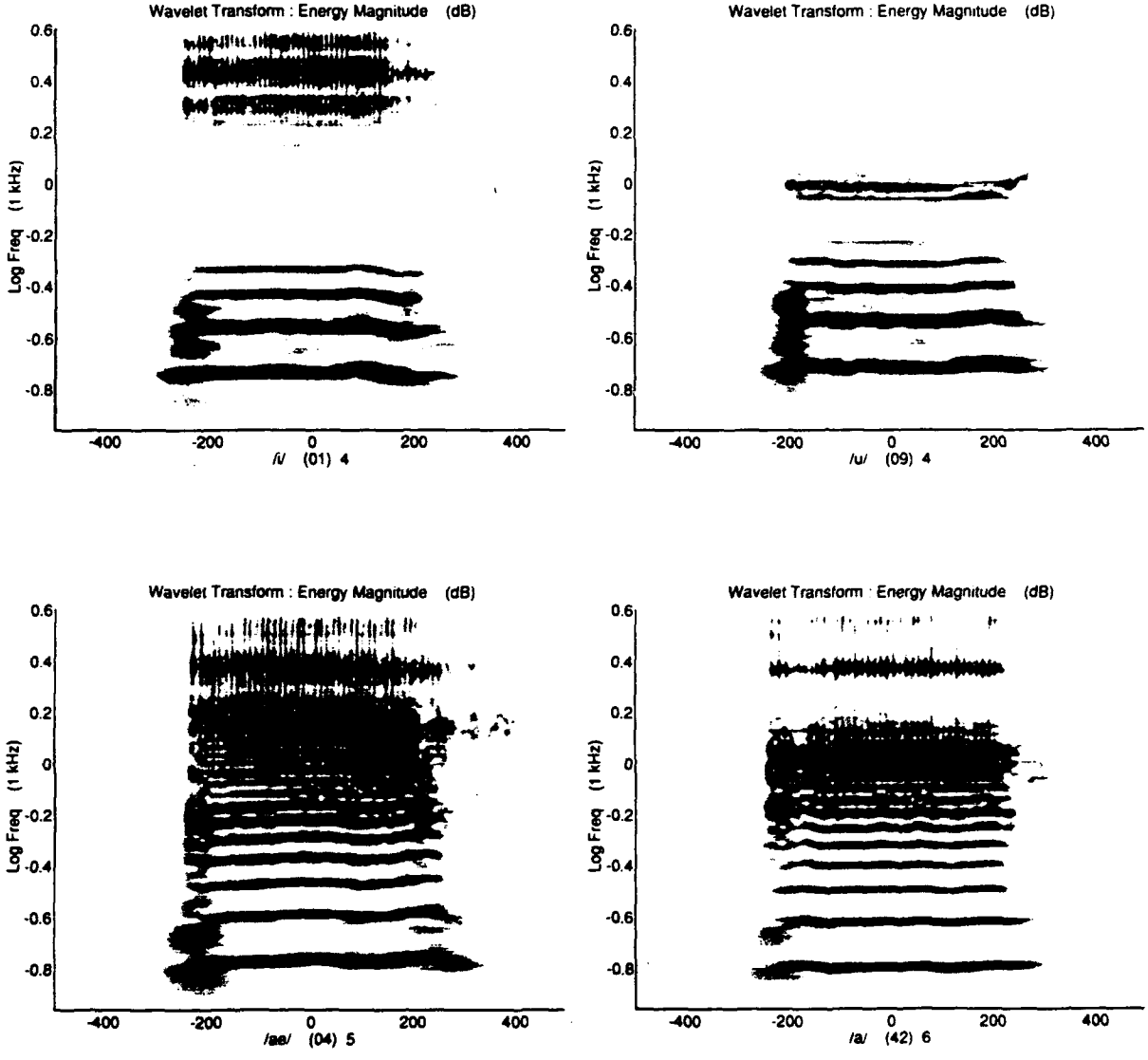


Figure 7.12 Wavelet Transforms of the 4 isolated vowels: /i/, /æ/, /ä/, /u/

Wavelet Transforms of the Gaussian WINDOWED vowels: /i/, /æ/, /ä/, /u/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

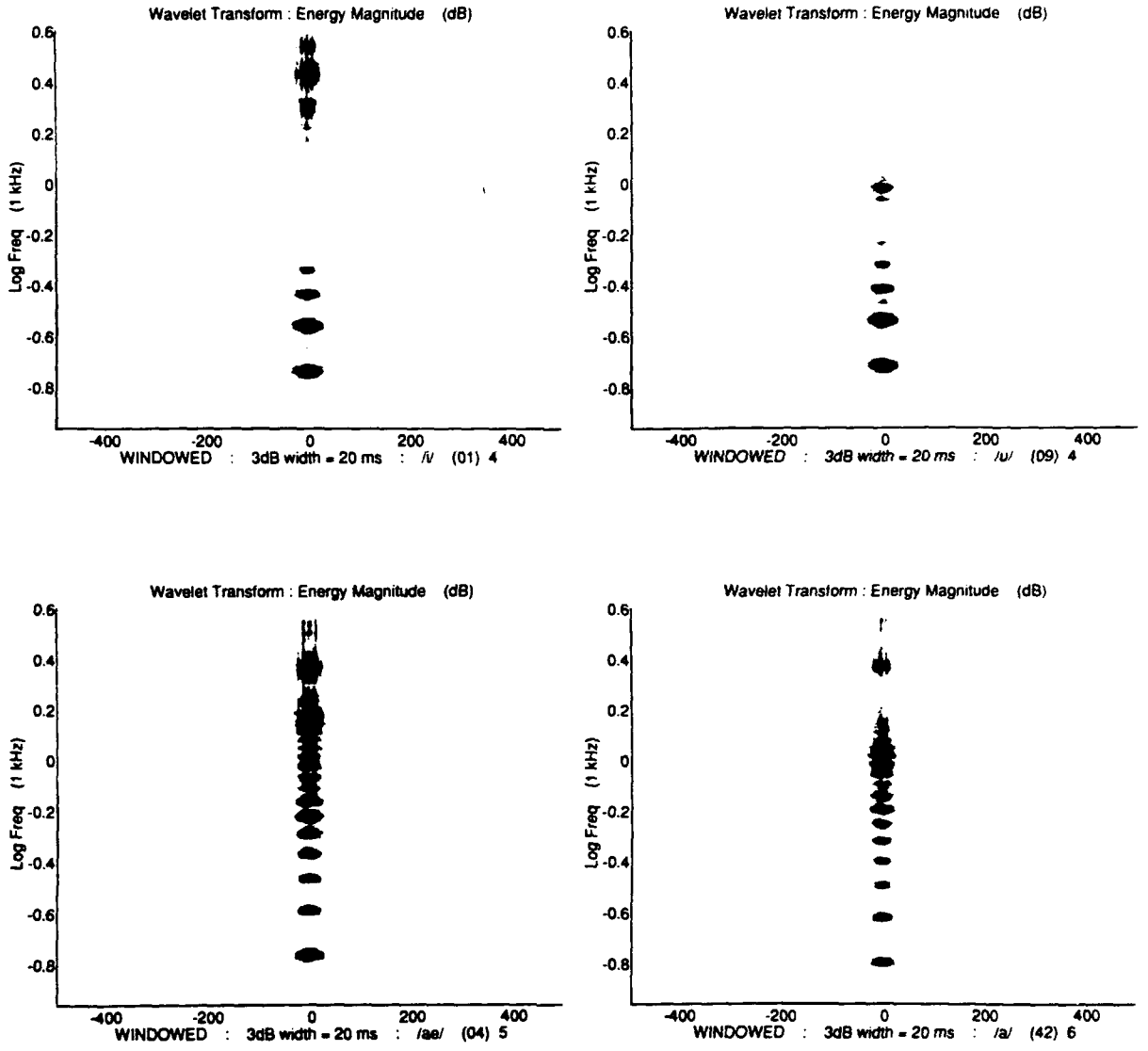


Figure 7.13 Wavelet Transforms of the Gaussian WINDOWED vowels: /i/, /æ/, /ä/, /u/

7.6 The Windowed COAR Results

The following figures present plots of the **[CÔAR]**(a,b) estimates which are calculated using the Gaussian windowed versions of the isolated vowels. In all cases, the same utterances and the same wavelet transforms are used as before in formulating the coarticulation channel estimates. The only exception, however, is the insertion of the time-windowing step on the wavelet transform of the isolated /V/ representation.

The modification generates a **[CÔAR]**(a,b) estimate for which the "control" articulation (the isolated vowel) has been windowed. As such, these "control windowed" versions of the **[CÔAR]**(a,b) estimates depict:

the cross wavelet correlation between 1) a complete CVC utterance

and 2) a windowed representation of the /V/.

Figure 7.14 contains the windowed **[CÔAR]**(a,b) estimates for each of the four vowels taken in their b/−/b context.

As a result of time-windowing, these **[CÔAR]**(a,b) plots occupy a reduced time-axis relative to the previous ones. The isolated vowel has been reduced in time, and so the total number of time points for this cross wavelet is now 453. There are still 40 points in scale. The interval between points on these grids is the same as before (the time points occur every 2.496 milliseconds).

The Figure 7.14 plots demonstrate that the windowing procedure is successful in yielding a **[CÔAR]**(a,b) distribution which is a dynamic function of time (b). Notice that the magnitudes at various scale values are synchronous in time. For example, the distribution for the [/i/, "beeb"] pair exhibits an abrupt initiation at time −100 ms along

Channel Estimate: WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ b/V/b

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

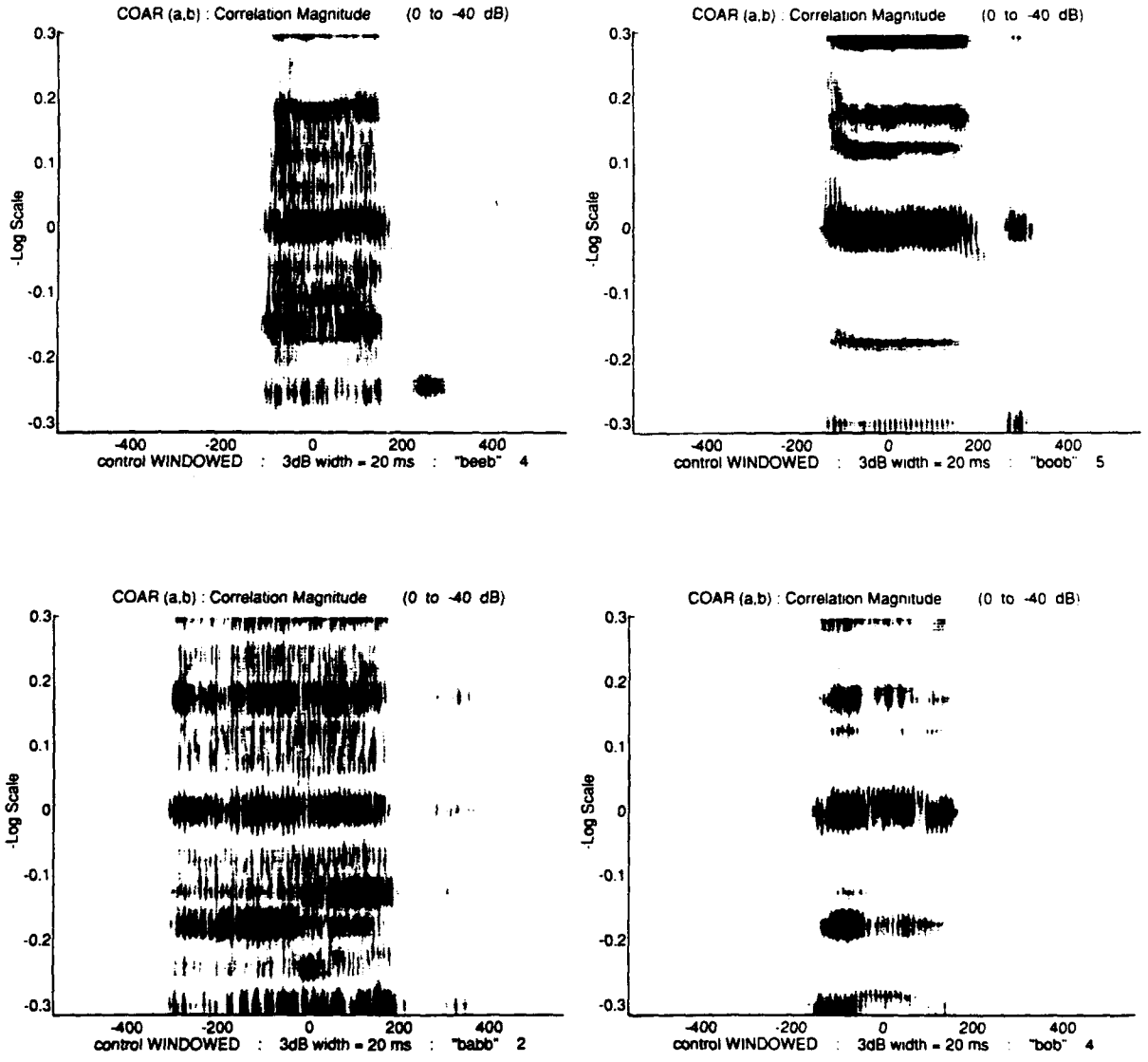


Figure 7.14 Channel Estimate:
WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ b/V/b

every scale value. This onset is associated with the initial /b/ burst in "beeb". In addition, a vertical white stripe appears between times +175 and +225 ms. This void is associated with the voicing stop-gap for the final /b/.

Similarly, (for other vowels) the initial stop burst of /b/ appears as an abrupt onset in the magnitude of the windowed **[CÔAR]**(a,b). The distribution rises synchronously at all scale values in the [/æ/, "babb"] plot, at time −300 ms. The same can be found in the [/u/, "boob"] plot, at time −175 ms.

The effect of the *final* /b/ burst, on the other hand, is especially evident in the plot of the [/u/, "boob"] correlation. At time +300 ms, there appears a vertical gray stripe. This highly localized, transient feature coincides with the release of the exploded (final) /b/ in "boob".

These attributes of the plots in Figure 7.14 attest to an improved time-variability in the modified **[CÔAR]**(a,b) estimate. None of the events cited here can be observed in the previous plots which were generated from the same sets of utterances (Figure 7.9). Furthermore, the good time-synchronism observed across scale values, and the local, transient character of various peaks indicate that the windowed **[CÔAR]**(a,b) responds *coherently* to dynamic gestures within these CVC articulations. More such evidence of time-variability appears in Figure 7.15, which shows the windowed **[CÔAR]**(a,b) estimates for the vowels in their *nasal* stop (m/−/m) context.

Apart from time-synchronism, there is another important element of time-variation to be observed in the windowed **[CÔAR]**(a,b) functions. The distributions shown in Figure 7.15 exhibit *transitions* between successive phones. Notice, in the cross between

Channel Estimate: WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ m/V/m

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

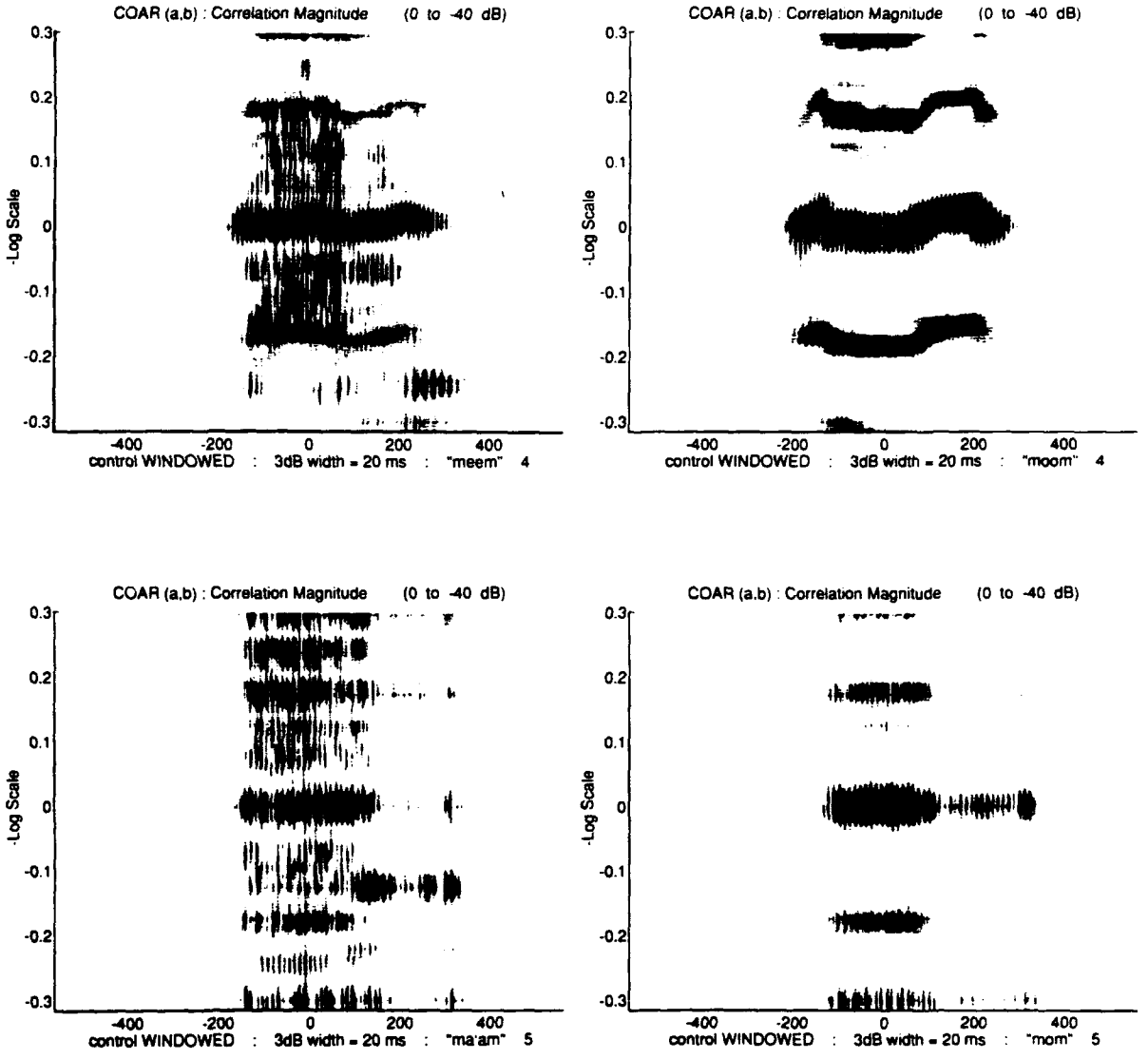


Figure 7.15 Channel Estimate:
WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ m/V/m

[/i/, "meem"], that the location in scale of the upper ridge ($-\log$ scale ≈ 0.2) undergoes some *displacement* with respect to time. Such ridge displacements as a function of time are even more apparent in the plot of the [/u/, "moom"] pair. The ridges in the latter plot are *continuously* maintained throughout "moom," yet, there is a transitional scale displacement out of /m/ and into /u/, at time -200 ms. Another scale displacement (out of /u/, into /m/) registers continuously over the interval between $+100$ and $+250$ ms.

Dynamic transitions between consonants and vowels are manifested not only in the form of scale displacements, however. The plots of Figure 7.15 exhibit clear variations in *magnitude*, attributable to the boundary regions between /V/ and (either) /m/. For example, the central ridge peak of the [/æ/, "ma'am"] pair ($-\log$ scale $= 0$) becomes severely attenuated at the onset of the final /m/ closure, yet, it is not completely muted. In contrast, a lower ridge (at $-\log$ scale $= -0.1$) experiences a dramatic *growth* at the time of closure for /m/ (time $+150$ ms).

Likewise, the plot of the [/a/, "mom"] pair (in the same figure) exhibits changes in ridge magnitude which are indicative of a vowel/consonant transition. The closure of the final /m/ is observable from the attenuation of all of the ridges (time $+100$ ms). Yet, the central ridge ($-\log$ scale ≈ 0) is sustained over the course of the closed /m/ voicing (time $+100$ to $+400$ ms). The magnitude of that ridge increases again momentarily (at time $+400$ ms), indicating the release of the final (exploded) /m/.

Compare these time-domain events (shown for the [/a/, "mom" 5] pair in Figure 7.15) with their counterparts observable in *another* plot: the *wavelet transform* plot of the CVC utterance ("mom" 5 ; Figure 7.6). Notice that the initial /m/ burst, final closure,

and final /m/ release are each visible from that spectral representation at the precise times cited in the windowed **[CÔAR]**(a,b) representation.

In summary, the time-variations in a windowed **[CÔAR]**(a,b) function are capable of depicting transient elements in a CVC articulation. This time variability can be manifested through fluctuations in ridge *scale location* (direction) or ridge *magnitude* (darkness). By direct comparison with a spectral (Morlet wavelet transform) representation of the lone CVC utterance, time-domain landmarks in the windowed **[CÔAR]**(a,b) are shown to be true and reliable indicators of real articulatory events. Furthermore, the configuration and continuity of these landmarks indicate that such variations are legitimate artifacts of the *coarticulation* occurring in the boundary region between consonant and vowel.

From a signal-processing point-of-view, the control windowed version of the **[CÔAR]**(a,b) estimate has substantially better *time resolution* than that of the original, unmodified version. It is therefore a superior representation of the coarticulation channel estimate. Note that in the initial, theoretical statement of the model, the effective time-resolution of the control vowel was not taken into consideration.

Yet, because the proposed theoretical model is designed to analyze CVC coarticulation via time and scale parameters, it should be responsive to the dynamic properties of that coarticulation *specifically within the domain of its time parameter* (b). The windowed version of the **[CÔAR]**(a,b), through its time-variability, manifests some of those dynamic properties. It thus leads to results which are more consistent with what is expected from the model. Through the remainder of this thesis, therefore, time

windowing the wavelet transform of the isolated vowel will be considered part of the standard procedure for calculating **[CÔAR]**(a,b) estimates.

7.7 Performance of the Windowed COAR for the Vowel /u/

In the previous section, it is shown that the windowed **[CÔAR]**(a,b) yields a superior estimate of the coarticulation channel function, in comparison to the unmodified version. Among the *windowed* **[CÔAR]**(a,b) estimates obtained in the study, however, it appears that those calculated for the vowel /u/ yield the clearest manifestation of consonant/vowel transitions.

Consider the plot of the [/u/, "boob"] cross presented in Figure 7.14. The transitions of the vowel to and from the adjacent consonants are evident from the vertical sweeps of the vowel's ridges at times −150 ms and +200 ms. The ridges, as sustained through the medial portion of the vowel, are swept upwards in scale at precisely the time of the initial stop-burst, and are swept downwards in scale at the time of the final stop-burst. Observe, more importantly, that each of the ridges is maintained *continuously* throughout these transitions. In other words, the same ridge responds (in different ways) over the course of the entire CVC articulation. This quality of ridge continuity from /C/ to /V/ to /C/ is an indication that the ridge's "trajectory" in the time-scale plane is specifically attributable to coarticulation. *The ridge trajectory is the model's response to the acoustic effect of CVC coarticulation on the vowel.*

The same observation can be made for the **[CÔAR]**(a,b) plot of the pair [/u/, "moom"] in Figure 7.15. A pattern of transitional ridge trajectories is easily identified

in the overall "S" shape of this plot. In this case, the initial and final transitions are sustained over longer time intervals (as compared with those of the "boob" plot, Figure 7.14). These longer transition intervals may be attributed to the longer closure period of the nasal stop.

Figure 7.16 contains the windowed **[CÔAR]**(a,b) estimates for the vowels embedded in their r/−/r context. Coarticulation for the /u/ vowel is depicted in the plot [/u/, "rure"]. A transition from the initial /r/ release to the medial portion of /u/ is observed (from time −200 to −100 ms) in a series of ridges having a concave downward trajectory. As in previous figures, this vowel exhibits the stronger ridges and the more discernible ridge trajectories (displacements in scale as a function of time), as compared to the other three vowels.

In conclusion, the vowel /u/ (when crossed with any consonantal context) yields in the **[CÔAR]**(a,b) function a series of horizontal ridges which exhibit distinct *directional* changes within the (a,b) plane. Some of the ridges turn concave upwards, and others turn concave downwards. Within a given [/u/, "C/u/C"] plot, however, the ridges are typically tracked in parallel with one another; i.e., they turn in the same direction at the same time. This is true whether the ridges are located above or below the −log scale = 0 mark. It appears that these variations are associated with the consonantal transitions into and out of the vowel. In such cases, therefore, the windowed **[CÔAR]**(a,b) provides a representation of the vowel which is sensitive to the quality and magnitude of coarticulatory effects.

Channel Estimate: WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ r/V/r

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

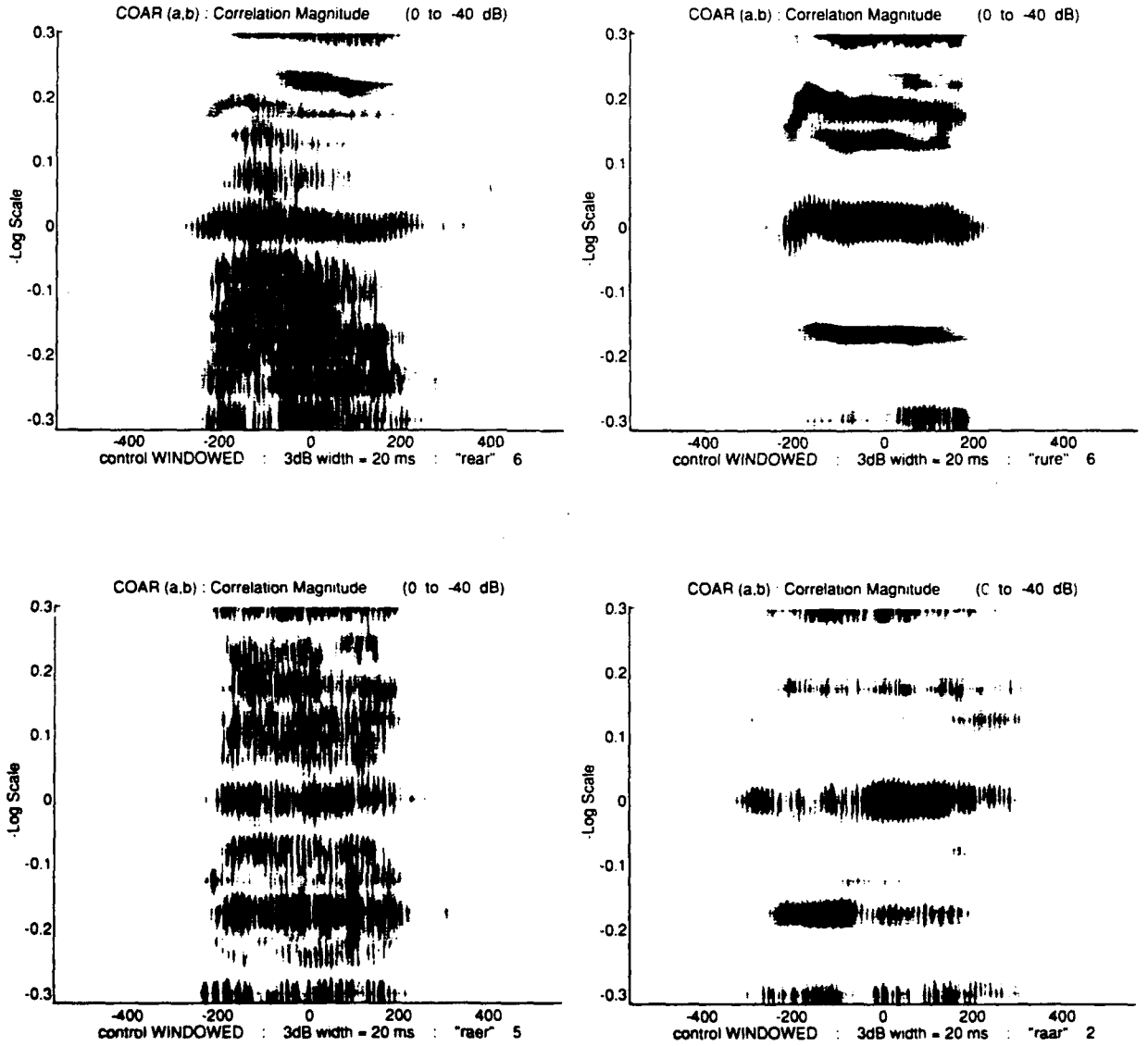


Figure 7.16 Channel Estimate:
WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ r/V/r

7.8 Some Observations of the COAR Formulated for r/V/r Context

As a group, the windowed **[CÔAR]**(a,b) functions calculated for the r/V/r context elicit some noteworthy characteristics. This section is dedicated to highlighting them.

Consider first the [/u/, "rure"] pair from Figure 7.16. Two *secondary* ridges are visible in this plot. They are located just above and below the principal ridge at −log scale = 0.2. These secondary ridges are not continuously maintained throughout the CVC, rather they are short-lived (relative to the other ridges of that plot). The lower of these secondary ridges (−log scale = 0.15) commences at time −150 ms. It ceases at about the time when the *upper* secondary ridge commences, 0 ms. The upper secondary ridge (−log scale = 0.25) then continues to time +200 ms, the approximate time of completion for the final /r/.

Turning to another vowel in the /r/ contextual set (Figure 7.16), a similar configuration of secondary ridges is observable from the [/i/, "rear"] pair. The secondary ridges in this plot are located at very much the same places in time and scale as those of the [/u/, "rure"] pair. No such ridges are apparent in either of the other two vowels of the /r/ set. Nevertheless, it is feasible that this particular configuration, with respect to the vowels /i/ and /u/, is a characteristic feature of initial /r/ coarticulation.

Finally, consider the plot of the [/ä/, "raar"] pair in Figure 7.16. Notice that the windowed **[CÔAR]**(a,b) estimate exhibits a number of ridge *magnitude* variations. These level fluctuations appear to coincide with the presence and absence of the /r/ retroflex articulation, in both the initial and final consonant positions.

The observations brought forth in this section are not variations of the scale *displacement* variety. They do, however, reflect specific coarticulatory effects. As with all of the windowed **[CÔAR]**(a,b) plots so far presented, these distributions depict an acoustic representation of the CVC utterance *in terms of* the isolated vowel; i.e., the representations use as a basis the signal associated with the isolated vowel.

Landmarks in the **[CÔAR]**(a,b) plots (systematic fluctuations in the distribution with respect to time) have been repeatedly documented in this and previous sections. These landmarks are variations of the implicit isolated vowel function. The **[CÔAR]**(a,b) distribution (and coarticulation channel model) therefore describe the CVC utterance from the perspective of those variations.

In many cases of **[CÔAR]**(a,b) plots, the landmark punctuates a time-interval which is closely associated with the consonant closure and/or burst. It is reasonable to conclude, therefore, that the **[CÔAR]**(a,b) function provides, in practice, *explicit information on perturbation of the vowel as a consequence of its close proximity to the consonant*. In these instances, the proposed coarticulation model functions in a manner consistent with its theoretical definition, that is, it provides a concise acoustic description of consonant-vowel-consonant coarticulation.

On the other hand, what has not been shown from these results is any clear pattern of **[CÔAR]**(a,b) behavior consistent with broad *phonetic* categories. For example, no particular pattern of ridge trajectories or landmarks can be observed for the CVC's consonantal place-of-articulation. Indeed, the gross similarities exhibited by the coarticulation channel function appear to be more correlated with the vowel class associated with the [/V/, C/V/C] pair than with its consonant class. Some limited patterns

in the fine structure of these plots have been found as indicative of the consonant class. However, none of the results have suggested that *all* vowels are subject to the *same* quality of perturbation with respect to a given consonant.

7.9 Evaluating the COAR Distribution with Help of the Spectrogram

The previous sections of this chapter have accomplished the following purposes: They have presented the calculated plots of the wavelet transform and cross wavelet functions, interpreted their mathematical meaning (in the context of what is already known about the speech), and established their validity in eliciting appropriate responses. For the purposes of further verifying the **[CÔAR]**(a,b) distributions (e.g., establishing their reproducibility for *repeated* utterances), a series of validation topics are addressed in the following chapter.

This final section of the current chapter provides a comparative evaluation of the coarticulation channel model. More **[CÔAR]**(a,b) distributions are presented for a variety of different consonantal contexts, but using one vowel which has been shown to yield the most favorable results: /u/. These results are displayed alongside the classical *spectrographic* representations of the (lone) CVC utterances. Each figure consists of a **[CÔAR]**(a,b) plot placed directly below a narrowband spectrogram of the associated CVC. In each case, the "time-shift" axis of the **[CÔAR]**(a,b) plot appears in direct alignment with the "elapsed-time" axis of the spectrogram, yielding an optimal means of visual comparison.

The purpose of this section is not to provide a definitive assessment of the cross wavelet function's performance in comparison to that of the classical spectrogram. Rather, the spectrograms are provided as a means of interpreting features within the **[CÔAR]**(a,b). For example, some **[CÔAR]**(a,b) landmarks are interpretable in the light of their counterpart features which normally appear in the spectrogram. Other **[CÔAR]**(a,b) landmarks, however, have no apparent manifestation in the spectrogram. In such cases, it is argued that the coarticulation channel model provides acoustic information about the CVC utterance which was not available traditionally.

Naturally, not all of the familiar articulatory features made visible by the spectrogram are necessarily conveyed in the **[CÔAR]**(a,b) representation. This is because the **[CÔAR]**(a,b) representation is in no sense an equivalent representation. To the contrary, many familiar spectrographic features become *de*-emphasized in the **[CÔAR]**(a,b) representation. It is not the purpose of this section, therefore, to account for all that is known about CVC articulations in the context of **[CÔAR]**(a,b) plots.

Figure 7.17 shows the spectrogram of the utterance "dude," together with the windowed **[CÔAR]**(a,b) plot of the [/u/, "dude"] utterance pair. The spectrogram is a short-time, fast-Fourier transform of the sampled signal. The size of the FFT is 1000. The size yields, in conjunction with the 31.25 kHz sampling rate, an effective bin spacing of 31.25 Hz. The time-window, however, is Hanning, with an effective analysis bandwidth of 45 Hz. The spectrogram is evaluated at regular time-shift intervals of 5 ms. This combination results in an adjacent-window overlap of 84%.

The gray-scale of the spectrogram shows the magnitude of the transform measured in dB. The darker areas of the plot indicate areas of higher magnitude, with a range

Narrowband Spectrogram: "dude" || Windowed CÔAR(a,b): [/u/, "dude"]

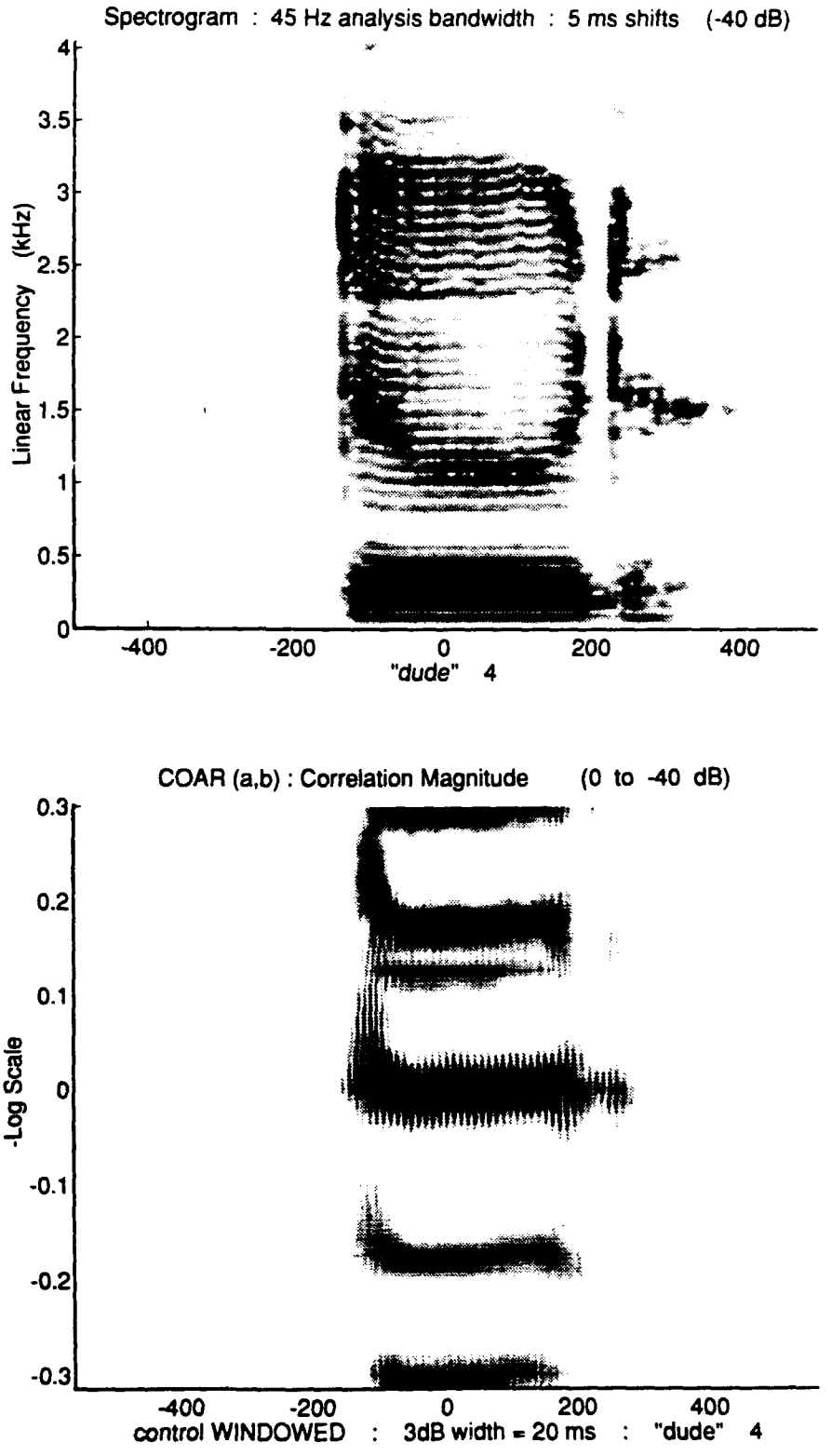


Figure 7.17

from 0 to −40 dB. The vertical axis is linear frequency measured in kiloHertz. The horizontal axis represents elapsed-time, measured in milliseconds.

The windowed **[CÔAR]**(*a*,*b*) plot appears in the lower portion of Figure 7.17. The physical attributes of this plot are the same as those presented previously. For the purposes of the figure presentation, however, the overall dimension of these plots has increased by about 50%.

The spectrogram clearly shows some elements of coarticulation occurring in the "dude" utterance. The coarticulation of /u/ attributable to the initial /d/ is particularly visible. During the medial portion of the vowel, from time −50 to +150 ms, the *F2* formant appears at about 1.0 kHz. From its transition out of the initial /d/, this formant has been dramatically stretched downwards in frequency. Between times −100 and −50 ms, *F2* falls about 500 Hz.

This [d/u] transition is also clearly visible in the **[CÔAR]**(*a*,*b*) plot. Notice the "swept ridge trajectories" exhibited over the same time intervals cited in the spectrogram (−100 to −50 ms). It is evident that these trajectories are a response to the perturbation of a vowel in transition. That is not to say, however, that this **[CÔAR]**(*a*,*b*) landmark represents an *F2* shift. Formant frequencies are measured in the spectral domain, which is *not* a physical dimension of the **[CÔAR]**(*a*,*b*). The ridge trajectories of the **[CÔAR]**(*a*,*b*) plot are indications of how a transition from /d/ to /u/ can undertake a *scaling* operation in the target vowel.

In the case of this consonant/vowel transition, therefore, both representations show a perturbation of the vowel /u/. In either case, the source of the perturbation is

clear (the vowel's close proximity to /d/). The physical *equivalence* of these features, however, has not been established.

Nevertheless, the ridge trajectories appearing in the **[CÔAR]**(*a*,*b*) plot can be interpreted independently of the spectrogram. For the [d/u] transition under consideration here, the trajectories indicate that the "vowel" is maintained continuously from its target value (at time −100 ms) backwards in time to the point-of-release of the initial /d/ (time −50 ms). In other words, beginning from the time of consonantal release, the acoustic structure of the vowel (or some modified version of it) is *present*. The initial vowel structure is then modified smoothly and continuously until the time that it reaches its target form (at time −100 ms). In short, this **[CÔAR]**(*a*,*b*) plot is evidence that the initial consonant contains the acoustic structure of a modified vowel.

Consider some additional examples of coarticulatory transitions. In the following cases, it will be shown that the alternative representations generate somewhat different responses. Figure 7.18 contains a spectrogram of the utterance "goog" alongside the **[CÔAR]**(*a*,*b*) plot of the utterance pair [/u/, "goog"]. Notice in the spectrogram that the initial consonant /g/ draws only a modest frequency-shift from *F2*. The formant varies in frequency about the distance of one harmonic (at time −100 ms). The cross wavelet plot, however, indicates a much more substantial and definitive coarticulatory effect at that time. The severity of this ridge trajectory appears to be of the same order as that observed for the initial /d/ of "dude" (Figure 7.17).

Figure 7.19 shows a spectrogram of the utterance "boob" alongside the **[CÔAR]**(*a*,*b*) plot of the [/u/, "boob"] pair. As in the case for "goog," the spectrographic representation reveals little or none of the formant frequency shifts which

Narrowband Spectrogram: "goog" || Windowed CÔAR(a,b): [/u/, "goog"]

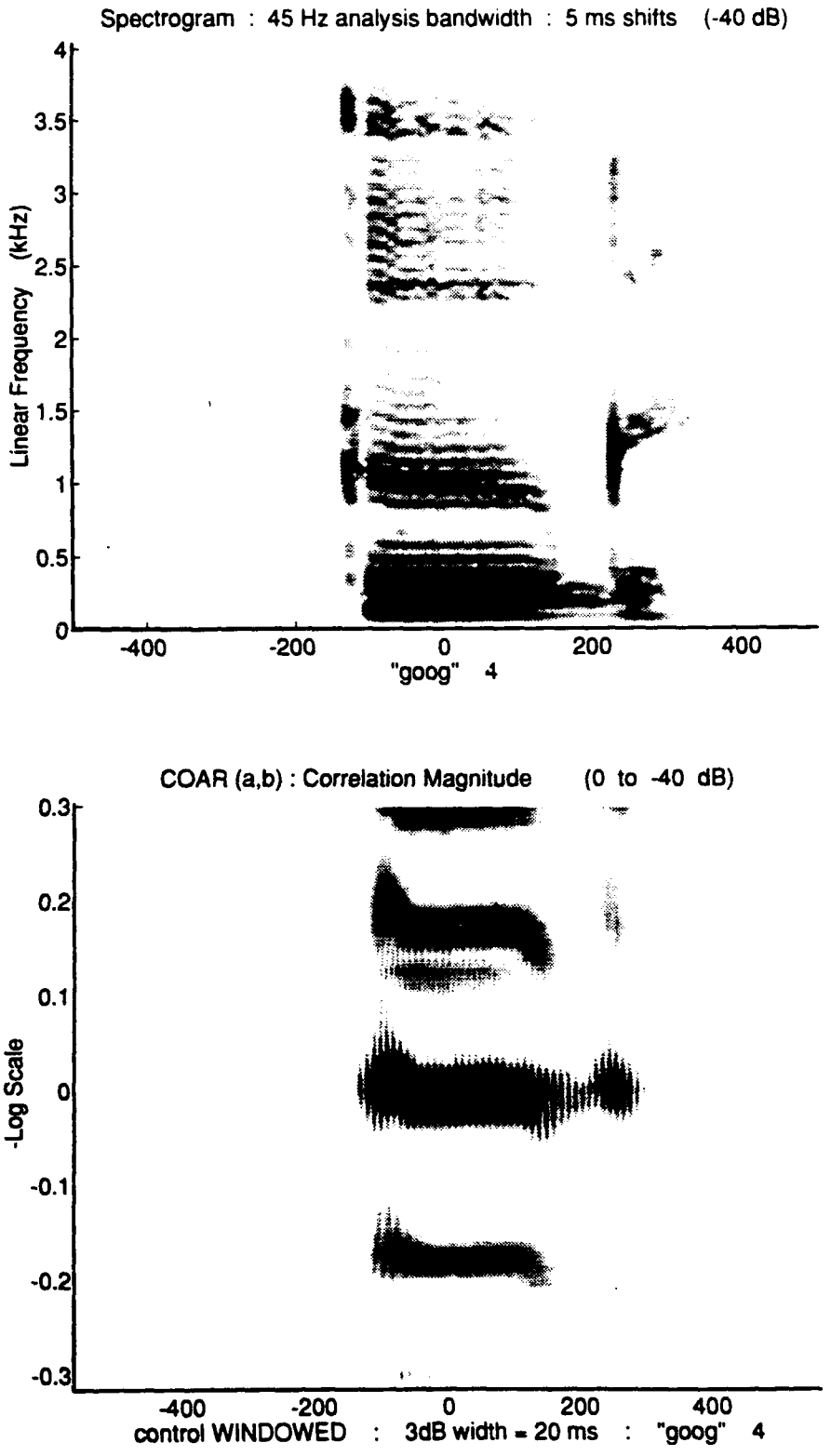


Figure 7.18

Narrowband Spectrogram: "boob" || Windowed CÔAR(a,b): [/u/, "boob"]

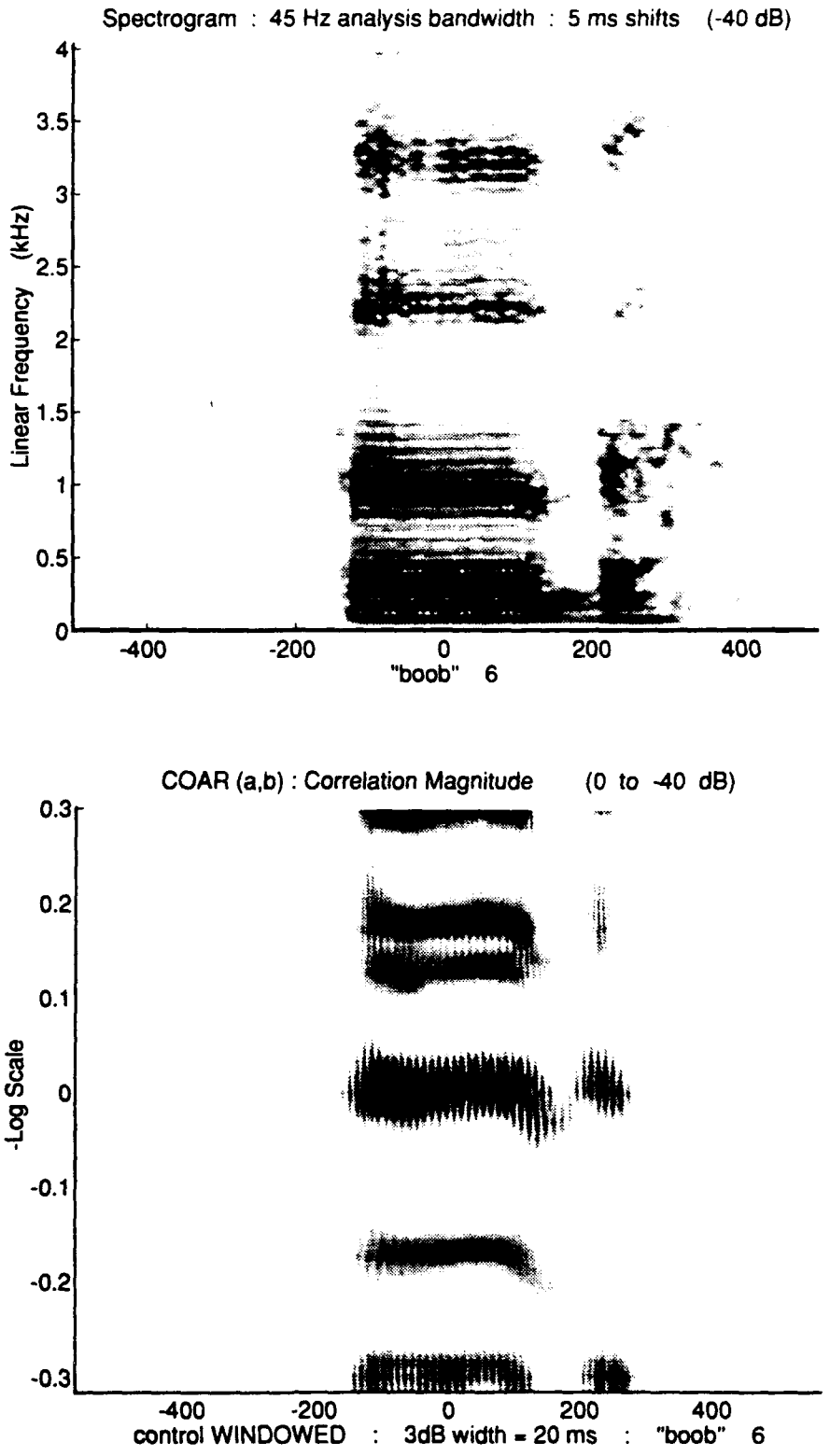


Figure 7.19

would normally indicate the presence of heavy coarticulation. This appears equally true for either of the initial or final consonant transitions. The cross wavelet plot, however, exhibits trajectory sweeping at the initial /b/, and shows more ridge bending at the final /b/. Observe, in particular, the winding trajectory of the final transition (from time +100 ms to +250 ms). The central ridge is roughly sustained throughout the entire duration of the final closure, stop-gap, and release.

These examples indicate that the coarticulation channel function is sometimes capable of delineating vowel perturbation effects with greater sensitivity, in comparison to the spectrogram. This greater sensitivity response is manifested by way of *greater scale*-dimension displacements and *longer duration* displacements. In other words, the vowel perturbation manifested by the **[CÔAR]**(a,b) function encompasses more remote scale-locations and more remote time-locations than does the spectrogram (in these particular cases).

One possible explanation for the spectrogram's relative deficiency in these cases is the poor frequency-resolution available in the low frequency region. As suggested previously, the *F1* formant in a spectrogram is often hardly distinguishable from the ridge of the fundamental. Any coarticulatory shifts in *F1* which *might* be occurring as a result of CVC coarticulation are not likely to show clearly in this region. Perhaps the magnitude of a frequency-shift exhibited on an *F1* formant is small in absolute terms, yet significant, considering the small value of *F1* itself. In contrast, the cross wavelet function, which occupies the *scale* dimension, inherently provides a relative measure for comparing such differences. Frequency shifts exhibited by the spectrogram, are (instead) evaluated in the cross wavelet function as adjustments in scale-factor.

The following examples show that the **[CÔAR]**(a,b) distribution can yield ridge trajectories through the closed interval of a nasal stop. Figure 7.20 presents the spectrogram and cross wavelet plots for the [/u/, "moom"] coarticulation category. Notice from this figure that the nasal *F1* attenuation so pronounced in the spectrogram (frequency 500 Hz and time +50 to +200 ms) is also detectable in the **[CÔAR]**(a,b) plot. An overall decrease in correlation magnitude is shown in the latter plot over precisely the same time interval; though, some cross wavelet ridges appear to be more attenuated than others. This does *not* suggest, however, that the magnitude fluctuations occurring over that interval of the **[CÔAR]**(a,b) can always be interpreted as nasal attenuations.

What is observable from the **[CÔAR]**(a,b) plot of Figure 7.20 are a series of ridge trajectories which extend back from the vowel to the initiation of the closed /m/ voicing. It is clear from the spectrogram that voicing for the initial /m/ begins at time −300 ms (100 Hz frequency location). Notice, however, the presence of **[CÔAR]**(a,b) ridge trajectories at that time. These trajectories are sustained from time −300 ms to a later time which apparently marks the release of the nasal stop, −200 ms.

A similar set of "pre-release" ridges are observable in the plots calculated for the /n/ version of the nasal. Figure 7.21 shows the spectrogram of "noon" along with the **[CÔAR]**(a,b) measured for the [/u/, "noon"] pair. On the interval from time −250 to −150 ms, the initial /n/ is in a state of closure. This is accompanied by a pattern of ridge transitions leading into the medial portion of the vowel. Incidently, the spectrogram shown in this figure differs physically from the other spectrograms: It is evaluated at regular time-shift intervals of 10 ms (rather than 5 ms).

Narrowband Spectrogram: "moom" || Windowed CÔAR(a,b): [/u/, "moom"]

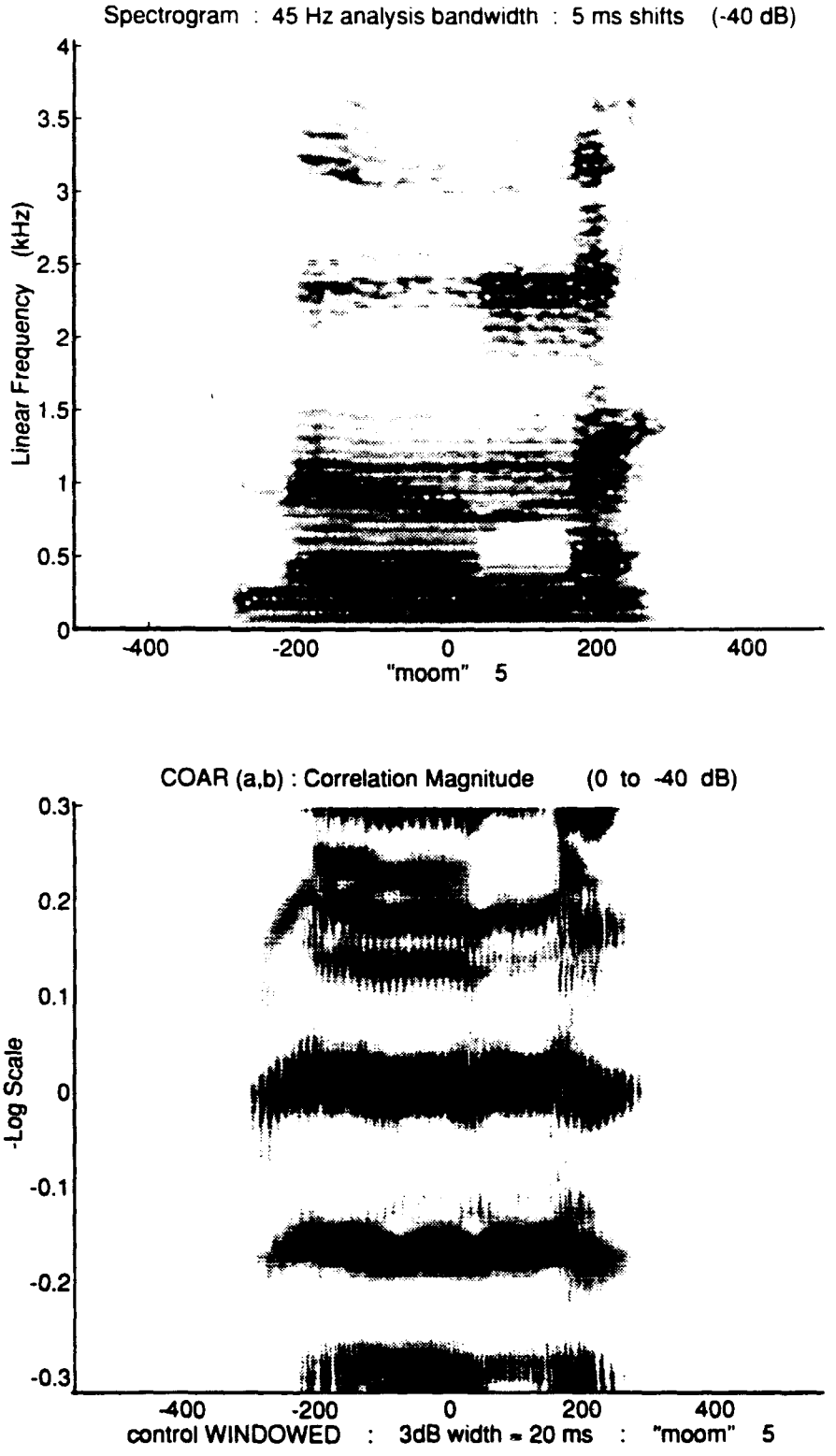


Figure 7.20

Narrowband Spectrogram: "noon" || Windowed CÔAR(a,b): [/u/, "noon"]

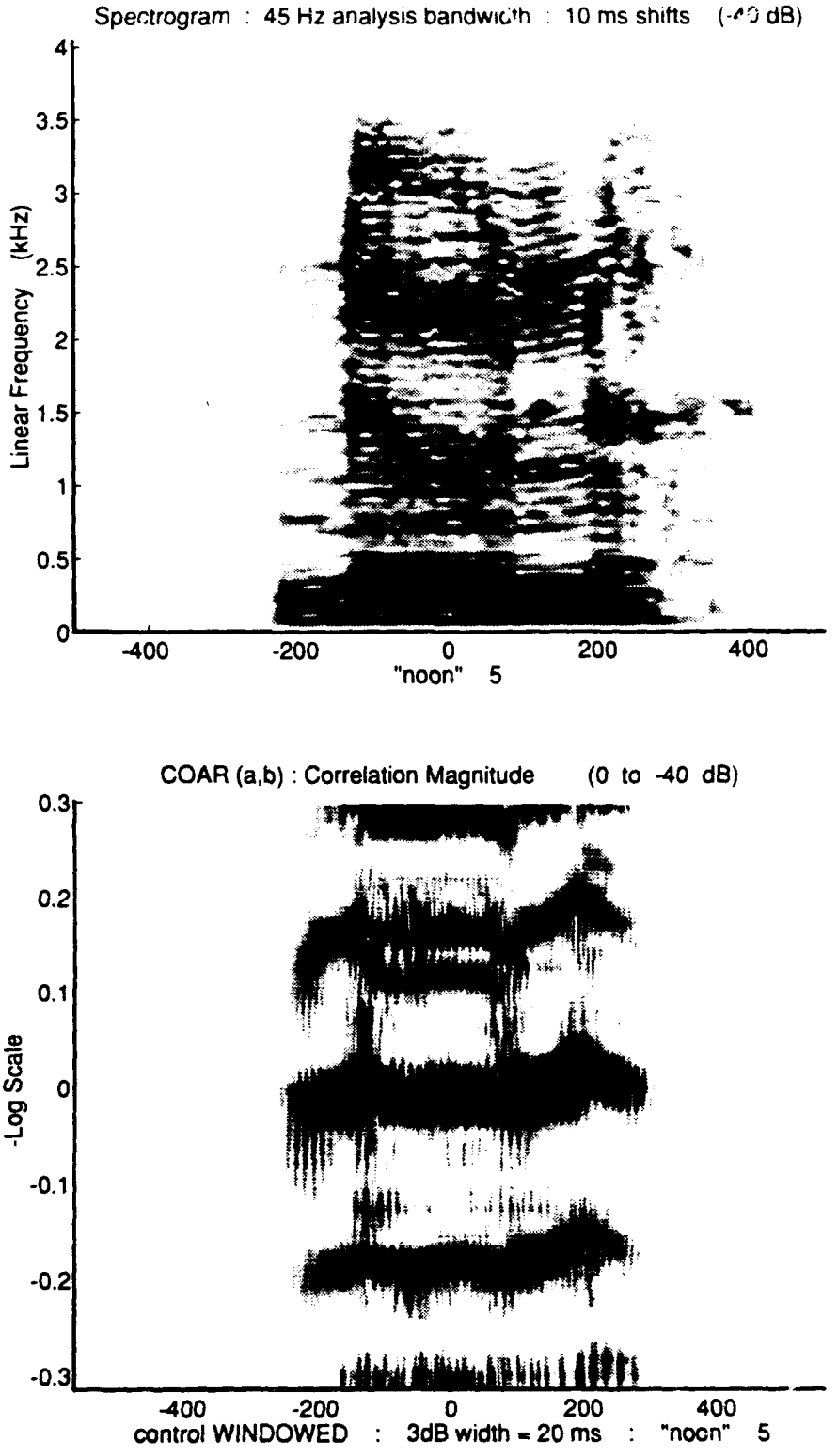


Figure 7.21

It is not clear whether the extended ridge trajectories observed in the cross wavelet plots for these nasal contexts should be interpreted as instances of vowel coarticulation. The true phonetic articulation associated with these "pre-release" intervals is most likely the sustained gesture of a *closed nasal consonant*. What appears in the **[CÔAR]**(*a*,*b*) over these intervals, therefore, may be the result of a "forced" attempt to interpret them as vowel-like articulations.

On the other hand, prior to the release of the nasal stop, the most prominent spectrographic feature is the fundamental frequency of voicing. It is feasible, therefore, that the cross wavelet ridge trajectories observed over these intervals stem from a pattern of minute deviations in the fundamental frequency of voicing. In other words, the **[CÔAR]**(*a*,*b*) ridge might respond as a correlation between the *F0* for the isolated vowel and the *F0* for the closed-stop portion of the CVC. The displaced trajectories are the result of slight transient deviations in fundamental frequency between the two. Notice that the time duration of the isolated vowel representation is not long enough to support fluctuations in *F0* over time (see Figure 7.13). However, fast *F0* fluctuations (too fast to be perceived as pitch deviations by a listener) might be enough to explain the scale displacements observed over these closed-stop intervals.

The final set of spectrographic comparisons appears in following set of figures. Figure 7.22 plots the spectrogram and cross wavelet distribution for the [/u/, "rure"] category. A noteworthy feature of the **[CÔAR]**(*a*,*b*) plot appearing in this figure is the re-appearance of the "secondary" ridges which were cited in previous cross wavelet plots under the context /r/. Refer to page 112 and Figure 7.16. The secondary ridges in the current example (Figure 7.22) are located at scale values 0.15 and 0.25. The

Narrowband Spectrogram: "rure" || Windowed CÔAR(a,b): [/u/, "rure"]

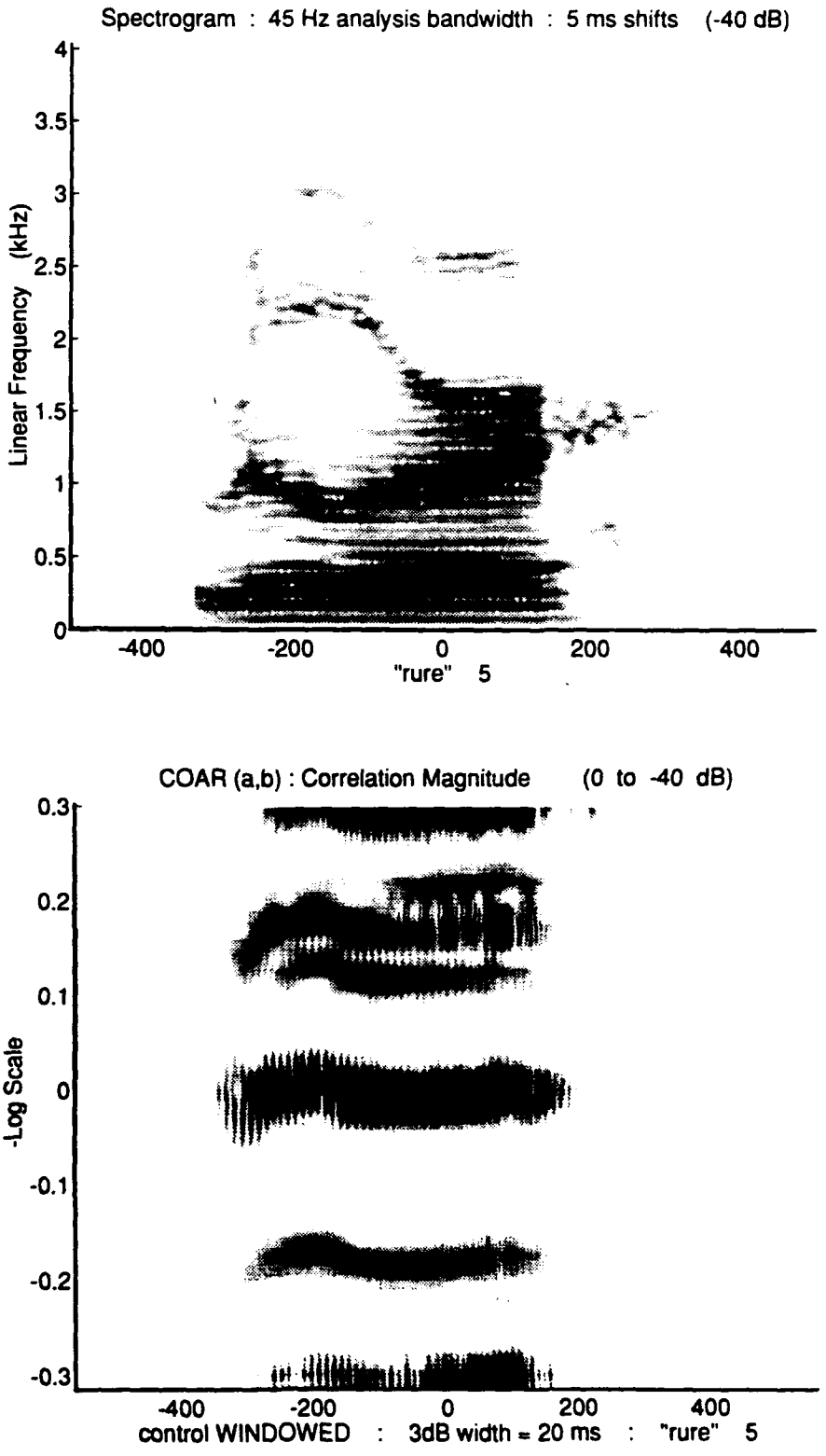


Figure 7.22

reproducibility of the ridge configuration in these examples suggests that this **[CÔAR]**(a,b) landmark might serve as an identifier for the retroflex/vowel transition.

Figure 7.23 likewise plots the spectrogram and cross wavelet distribution calculated for the [/u/, "lool"] category. Notice that the "lool" spectrogram shows a strong coarticulatory shift in the *F3* formant. Over the course of the vowel's medial portion, through to the final /l/ (−150 to +150 ms), *F3* rises from 2.3 to 2.7 kHz. However, no associated ridge displacements are apparent from the **[CÔAR]**(a,b) plot. In the case of the *initial* consonant, on the other hand, a strong transition does appear in the cross wavelet plot. This transition is marked by a set of swept ridge trajectories, oriented in time about the instant of the initial /l/ release (−150 ms).

7.10 Results Summary

A series of wavelet transform and cross wavelet transform calculations have been generated for a large subset of utterances obtained from the speech sample. The results of these calculations have been presented in the form of three-dimensional gray-scale plots. With regards to the question on how these new wavelet distributions should be interpreted, many observations have been made and a number of assertions have been contended. The assertions can be grouped into a few general categories, and these include: the effectiveness of the proposed model, the technique for optimizing the model, the acquisition of new information, and the disadvantages of the proposed model.

The first general assertion obtained from these data is that the proposed coarticulation model, when evaluated in practice, specifically provides a representation

Narrowband Spectrogram: "lool" || Windowed CÔAR(a,b): [/u/, "lool"]

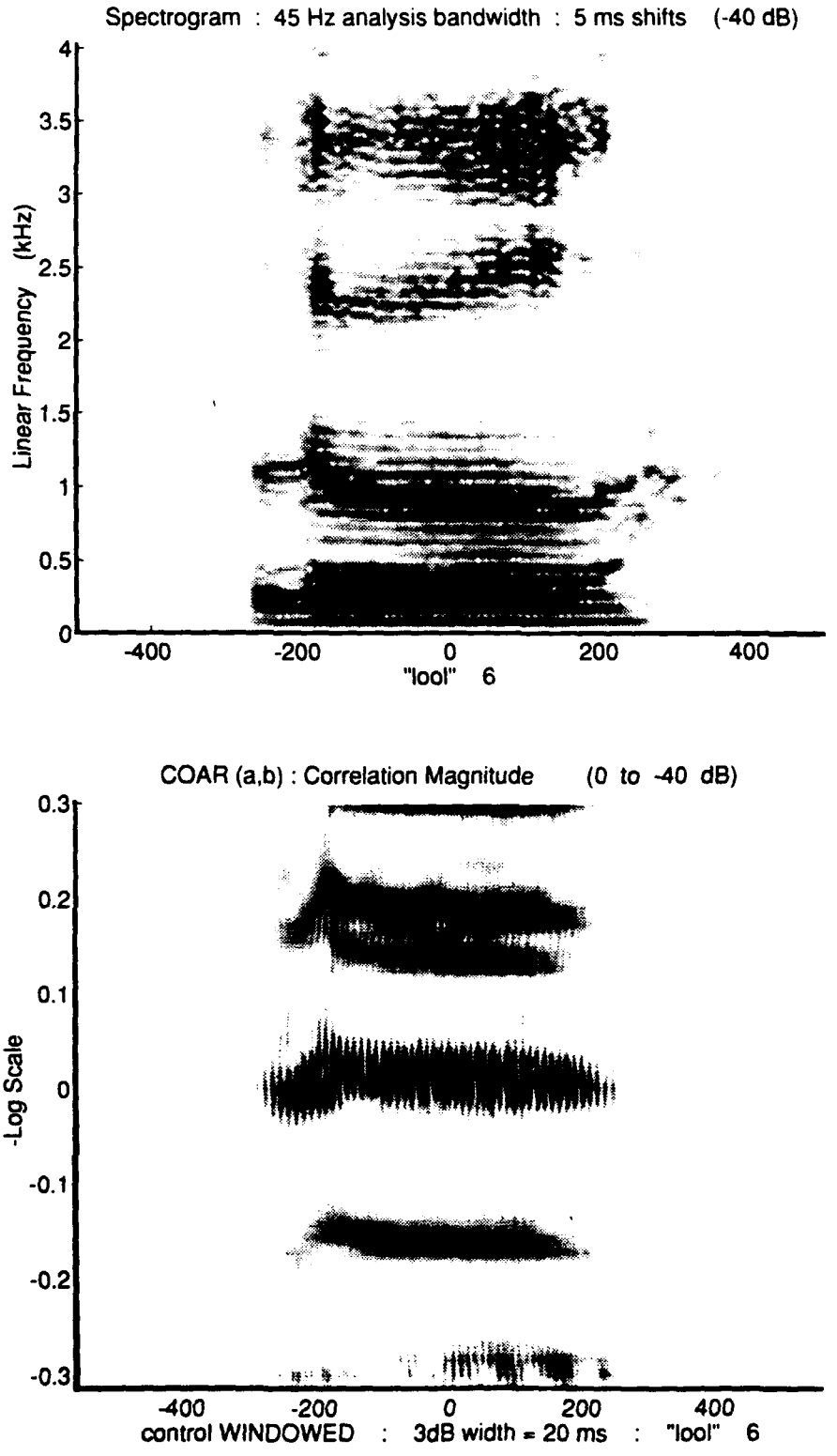


Figure 7.23

for acoustic effects in CVC coarticulation. The configuration and continuity of the various landmarks observed in the **[CÔAR]**(a,b) indicate its responsiveness to legitimate artifacts of coarticulation, in the boundary region between consonant and vowel. Taking into consideration the physical meaning of the cross wavelet channel characterization, these calculated distributions yield an explicit illustration of the vowel's perturbation as a consequence of its close proximity to the consonant.

The model performs better in the case of one vowel (/u/) than for the other vowels. The **[CÔAR]**(a,b) response for /u/ coarticulation is typically manifested in a grouping of ridge-displacement trajectories. These ridge trajectories convey some detail regarding the magnitude and duration of the coarticulatory effect.

The second general assertion drawn from this data pertains to a technical procedure necessary for deriving meaningful estimates. It was established that the windowing modification for the **[CÔAR]**(a,b) distribution results in a substantially improved time-resolution. The gain in time-resolution is achieved without incurring the signal distortions which would normally be incumbent on such windowing. It was also found that the cross wavelet representation, when treated to this modification, responded coherently to dynamic gestures within CVC articulations.

Another general contention formulated from close examination of data is that, in select cases, the coarticulation channel model provides information about CVC coarticulation which is not readily available from traditional methods of analysis. For example, the model was shown to be capable of delineating certain vowel perturbation effects with greater sensitivity, in comparison to the spectrogram. The greater sensitivity

response is manifested in terms of the boundary transitions which occupy longer duration intervals and larger spans of frequency/scale-displacement.

Apart from its sensitivity to transitional effects, the coarticulation channel was also shown to provide new evidence regarding the acoustic *structure* of the consonant/vowel transition. That is, the underlying structure of a vowel can be found (in a perturbed form) throughout the entire interval of a transition, up to and including the point of consonantal release.

In the case of the Morlet wavelet transforms, it was shown that, for certain events, superior time resolution can be achieved at *all* frequencies, through the use of a simple extrapolation method. The method "borrows" the good time-resolution at high frequencies, and extrapolates down to low frequencies, for the purposes of pinpointing the precise time-location of a consonantal impulsive burst. In addition, it was found that the Morlet wavelet transform provided good definition and separation between the fundamental-frequency voicing bar and the first formant peak. Such separation is not normally achieved by the spectrogram for high vowels.

The last general category of observations addresses the problems and disadvantages incurred in using the proposed model. One of these is the potential *loss* of meaningful information. It was shown on several occasions that many features appearing in the **[CÔAR]**(a,b) distribution are *not* readily interpretable in light of the classical spectrogram. Because the current and classical methods of analysis are not equivalent, correlations between their associated features cannot always be established.

Secondly, it was shown that, in the cases of nasal CVC utterances, leading ridge trajectories in the **[CÔAR]**(a,b) distribution are not interpretable as indications of

consonant/vowel coarticulation. In particular, there is a question as to what role a slight deviation in fundamental frequency might play in the formulation of these trajectories.

Finally, it has been shown that the coarticulation channel function yields a good discrimination between the classes of vowels employed in this study. However, no clear patterns in the **[CÔAR]**(*a*,*b*) distributions have emerged with respect to the *consonant* class or place of articulation. The results do not suggest, for example, that all vowels are subject to the same quality of perturbation, with respect to any given phonetic class or distinction.

Chapter 8

VALIDATION

This chapter performs a number of empirical tests on the results derived from the experimental study. The purpose of these tests is to verify that the calculated **[CÔAR]**(a,b) distributions are reasonable results for what should be expected from the coarticulation model. Naturally, there are no previous data of this type to provide an external reference or basis of comparison. Using some special cases of the present data, however, some *qualitative* comparisons can be made to help establish the validity of this body of calculations.

The chapter is composed of four such comparisons. The first addresses the technical implementation of the model with respect to the relationship between $z2$ and C/V/C. This test pertains specifically to the inclusion of the consonant portions of the CVC as part of the signal employed in the implementation. The second assesses the "self-similarity" of the vowels, through a close examination of each vowel's auto-ambiguity function. This is followed by an examination of the model in its null state. That is, what happens to the model when no consonants are spoken for *either* utterance? Finally, a comparison of results generated from *repeated* utterances is presented as a descriptive measure of the model's overall reproducibility.

8.1 Evaluating the Inclusion of Consonants in *z2*

The purpose of this section is to validate the processing technique used in implementation, whereby, consonants are not explicitly removed from *z2*, the signal associated with the CVC utterance. The technique and the reasons for its utilization were outlined previously in the experiment chapter, section 6.8. (In theory, the coarticulation model poses a correlation between two vowels only, the isolated vowel and the vowel portion of the CVC utterance. The removal of consonants from the CVC signal, however, constitutes phonemic *segmentation*, a procedure which can generate additional discrepancies and ambiguities.)

It is shown presently that, using the entire CVC utterance in the cross with /V/, the same results are yielded as when the consonants of the CVC *are* carefully removed from the *z2* waveform. The implications of this comparisor are that the employed technique yields a good approximation to the strict implementation which uses only vowels. The test is conducted using each of the four vowels in the context of d/−/d.

Figure 8.1 shows the wavelet transforms of the four vowels imbedded in their d/−/d context. Notice in these plots the visibility of the final exploded /d/. The plosive burst of each final /d/ is shown by an abrupt vertical striation. This vertical striation closely follows a white vertical stripe, the voicing gap. The stop burst associated with the *initial* /d/ is visible for two of the vowels, /u/ ("dude") and /ä/ ("dodd"). Bursts from the initial /d/ are similarly manifested by thin vertical stripes, occurring in "dude" at time −200 ms, and in "dodd" at time −250 ms.

Wavelet Transforms of the /d/ words: /did/, /dæd/, /däd/, /dud/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

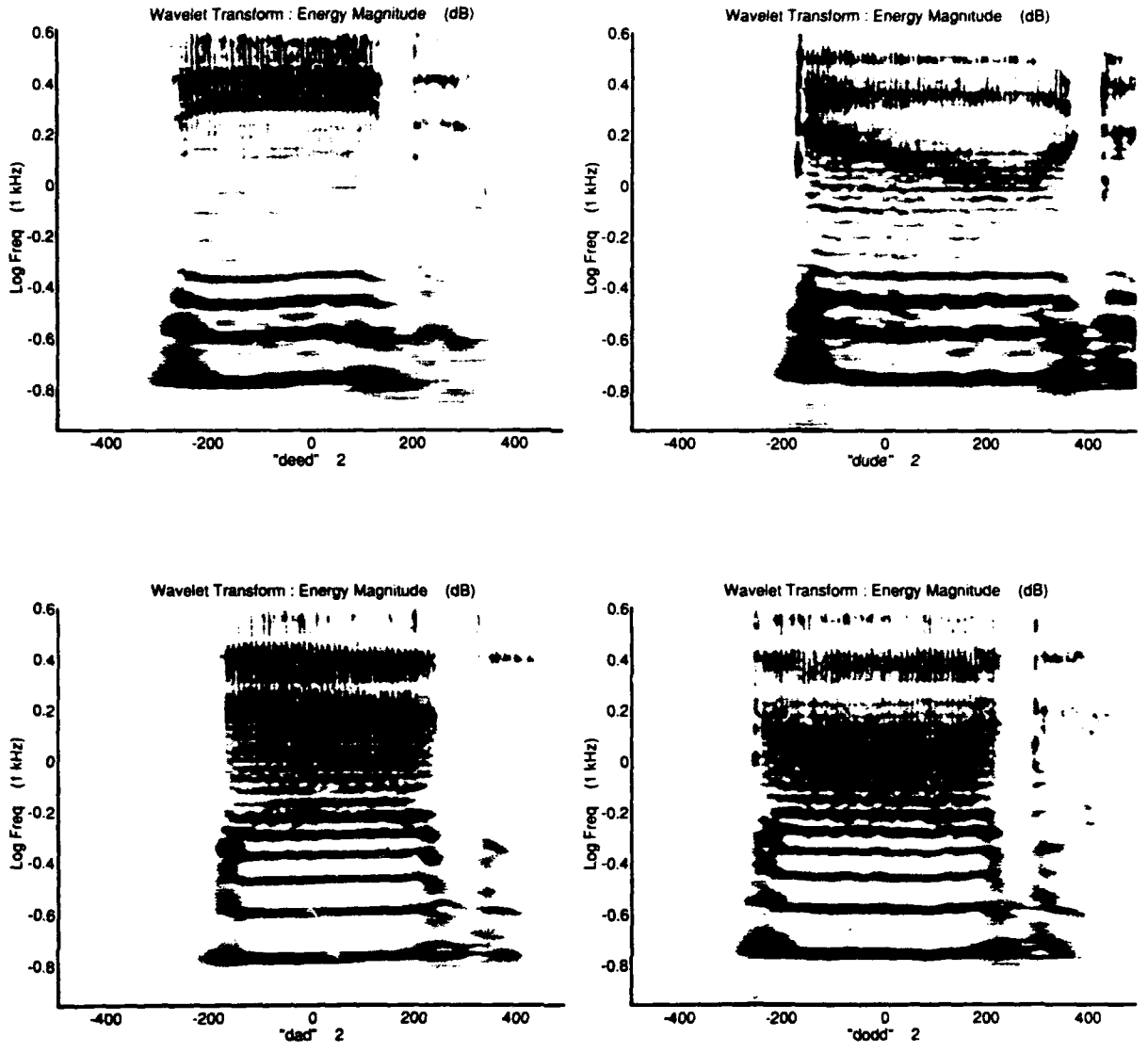


Figure 8.1 Wavelet Transforms of the /d/ words: /did/, /dæd/, /däd/, /dud/

Wavelet Transforms of CONSONANT CUT /d/ words: /did/, /dæd/, /däd/, /dud/

(0 to −40 dB) vs. Log Frequency (kHz) vs. Time (milliseconds)

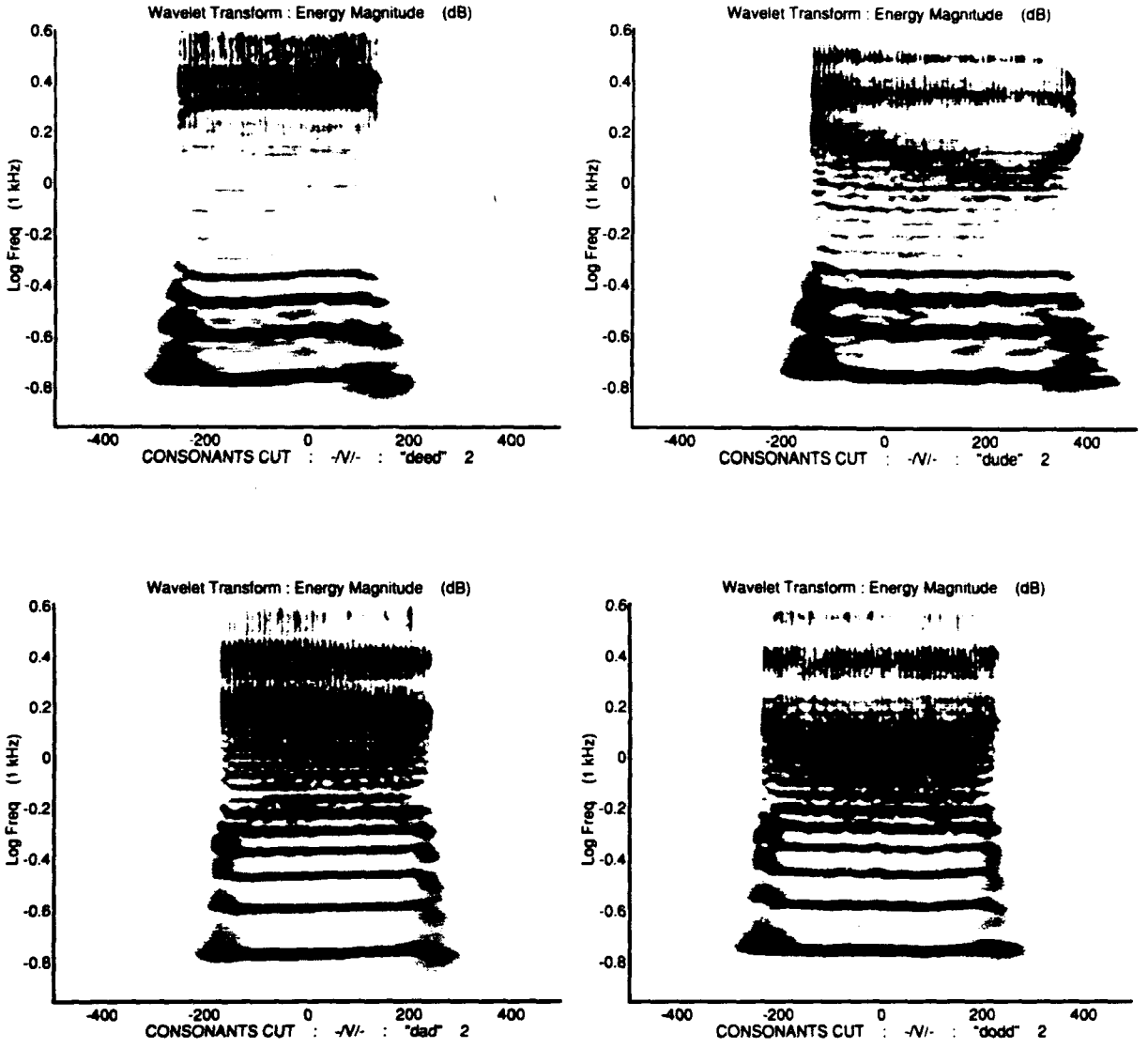


Figure 8.2 Wavelet Transforms of CONSONANT CUT /d/ words: /did/, /dæd/, /däd/, /dud/

Compare these plots with the plots in the next figure. Figure 8.2 presents the wavelet transforms calculated for the same four utterances, but this time the initial and final /d/ consonants were first cut from each waveform. More specifically, samples from the very beginning and very end of the signal were removed (set equal to zero). The choices of the segmentation boundaries (where the initial /d/ "ends" and where the final /d/ "begins") were determined from visual examination of the sampled waveform. Each segmentation boundary was also tested through audition of the resulting edited signal. Whether monitoring the waveform in a visual or auditory mode, it was found that (for the /d/ consonant) clear distinctions could be made between the location of the burst and the outside boundary of periodicity. Under this criterion, therefore, the burst portions were removed from the signal, and the periodic portion (middle) was retained. The Morlet wavelet transforms which appear in the figure were then implemented on the edited signals in the usual fashion.

Observe from Figure 8.2 the absence of the stop-burst landmarks (in the context of both the initial and final /d/ positions). Furthermore, comparing this figure with the previous Figure 8.1, notice the consistency between the (remaining) vowel portions. The plots of Figure 8.2 appear to be replicas of those from Figure 8.1, with the exception of the burst landmarks. This consistency between vowel pairs is anticipated. It testifies to the ability of the wavelet transform to discriminate in time. In other words, the removal of a consonant from time location t_1 has little or no effect on the vowel which appears at a *different* time location, t_2.

The following pair of figures is analogous to the previous pair. Figures 8.3 and 8.4 contain the windowed **[CÔAR]**(a,b) distributions which are associated with the

Channel Estimate: WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ d/V/d

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

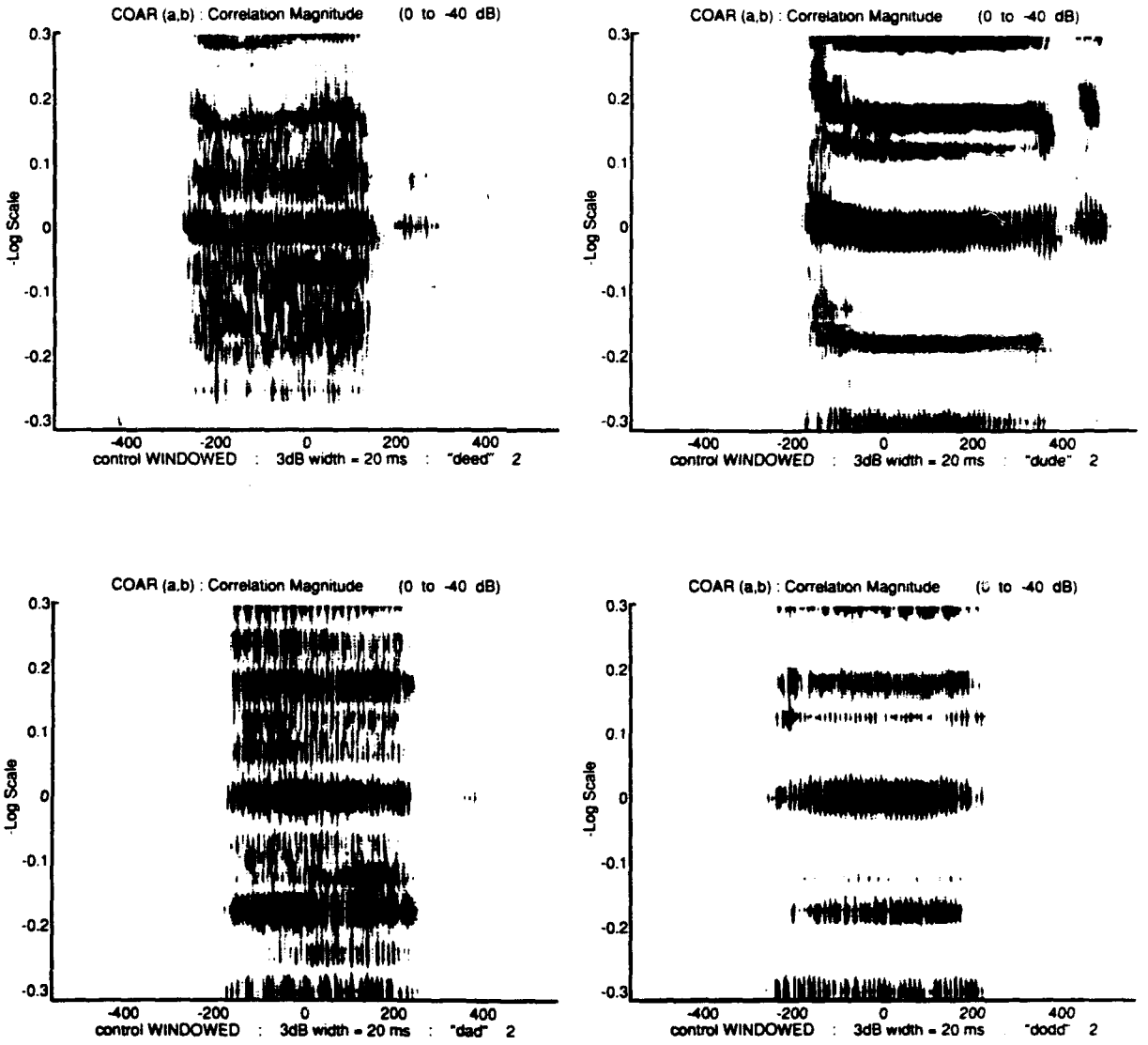


Figure 8.3 Channel Estimate:
WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ d/V/d

Channel Estimate: **WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ CONSONANTS CUT d/V/d**

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

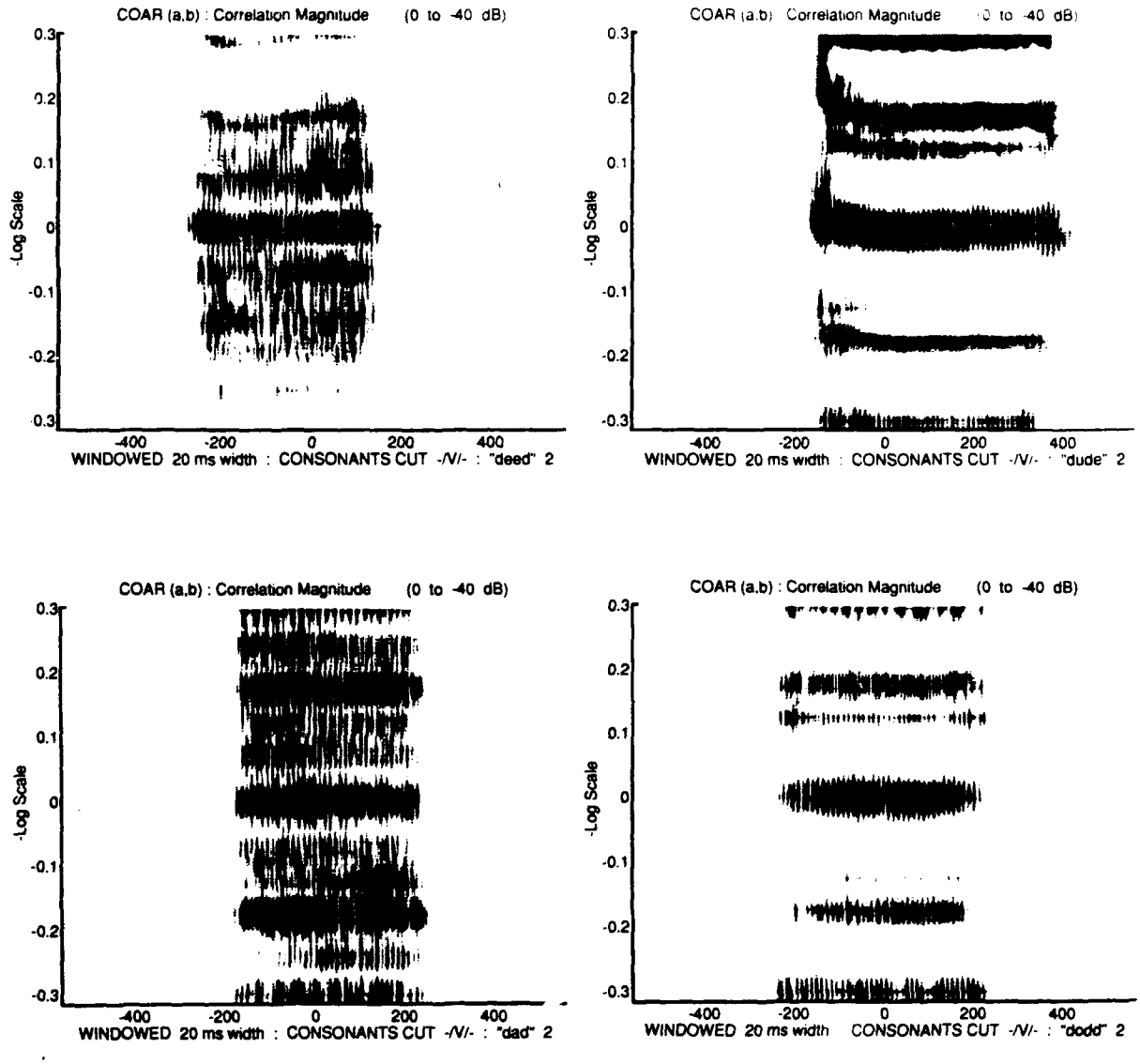


Figure 8.4 Channel Estimate:
WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ CONSONANT CUT d/V/d

earlier wavelet transforms. Figure 8.3 presents the cross wavelet plots for the vowels embedded their d/−/d context: [/i/, "deed"], [/æ/, "dad"], [/ä/, "dodd"], and [/u/, "dude"]. These cross wavelet distributions were calculated using each of the wavelet transforms appearing in Figure 8.1 as the CVC contingent. Likewise, Figure 8.4 contains the cross wavelet distributions calculated using the transforms of Figure 8.2 as the CVC contingent. In other words, Figure 8.4 presents the cross wavelet plots of the vowels embedded their d/−/d context, whereby, the *edited* (consonant-cut) versions of the CVC signals are employed.

The purpose behind these figure presentations is to show that the results (Figures 8.3 and 8.4) are quite comparable, except for the familiar stop-consonant landmarks which are present in Figure 8.3. The vowel portions of these windowed **[CÔAR]**(*a*,*b*) distributions are therefore *not* sensitive to the inclusion/omission of the consonants in the original CVC signal. Because the resultant **[CÔAR]**(*a*,*b*) distributions are primarily invariant, whether or not the initial and final consonants are explicitly removed, then the procedure which *includes* these consonants in processing is valid.

This comparison also gives evidence that the coarticulation channel should be viewed as a correlation between *vowels*, and not as an absolute analysis of the CVC utterance. Notice, in the case of the utterance pairs [/æ/, "dad"] and [/ä/, "dodd"], Figures 8.3 and 8.4 are virtually indistinguishable. Peaks in the **[CÔAR]**(*a*,*b*) distribution are registered only when the control utterance, /V/, is shown to be strongly similar to some portion of the CVC. Presumably, this can happen only during the vocalic, vowel-like portions of the CVC, to the exclusion of the purely consonantal ones.

8.2 The Auto-Ambiguity Functions of the Four Vowels

This section presents the auto-ambiguity functions of the isolated vowels. An auto-ambiguity function is the cross wavelet transform of a signal onto itself, or the wavelet transform of some function, using the same function as the mother wavelet.

As stated previously (page 93), the auto-ambiguity of an isolated vowel is a measure of the vowel's "self-similarity". (This applies whether or not the vowel representation has been windowed in the manner of section 7.5). Furthermore, it was shown in section 7.3 that the **[CÔAR]**(a,b) contains many aspects of the vowel's self-similarity. Simply because the same /V/ recurs in both the isolated and CVC utterances, the **[CÔAR]**(a,b) behaves, at certain time-locations, like an auto-ambiguity function.

The vowels' ambiguity functions, therefore, were calculated as a way to document a trivial case: the cross wavelet transform between a vowel and itself. Their plots are presented in Figure 8.5. Each plot shows the auto-ambiguity function of the vowel formulated from the windowed representation of that vowel. The following Figure 8.6 shows exactly the same calculations, with the exception that the range of time-points in each plot has been reduced from 1000 ms to 250 ms. This reduction in time range has the effect of time-wise "zooming" these presentations.

The plots illustrate concisely the patterns of self-similarity, associated with each vowel, which appear in most of the previous **[CÔAR]**(a,b) distributions. They also affirm the overall contrast *between* vowels. Consider that any differences between vowels shown in these figures are derived from differences in the vowels themselves.

Auto-Ambiguity function: [WINDOWED /V/, WINDOWED /V/]

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

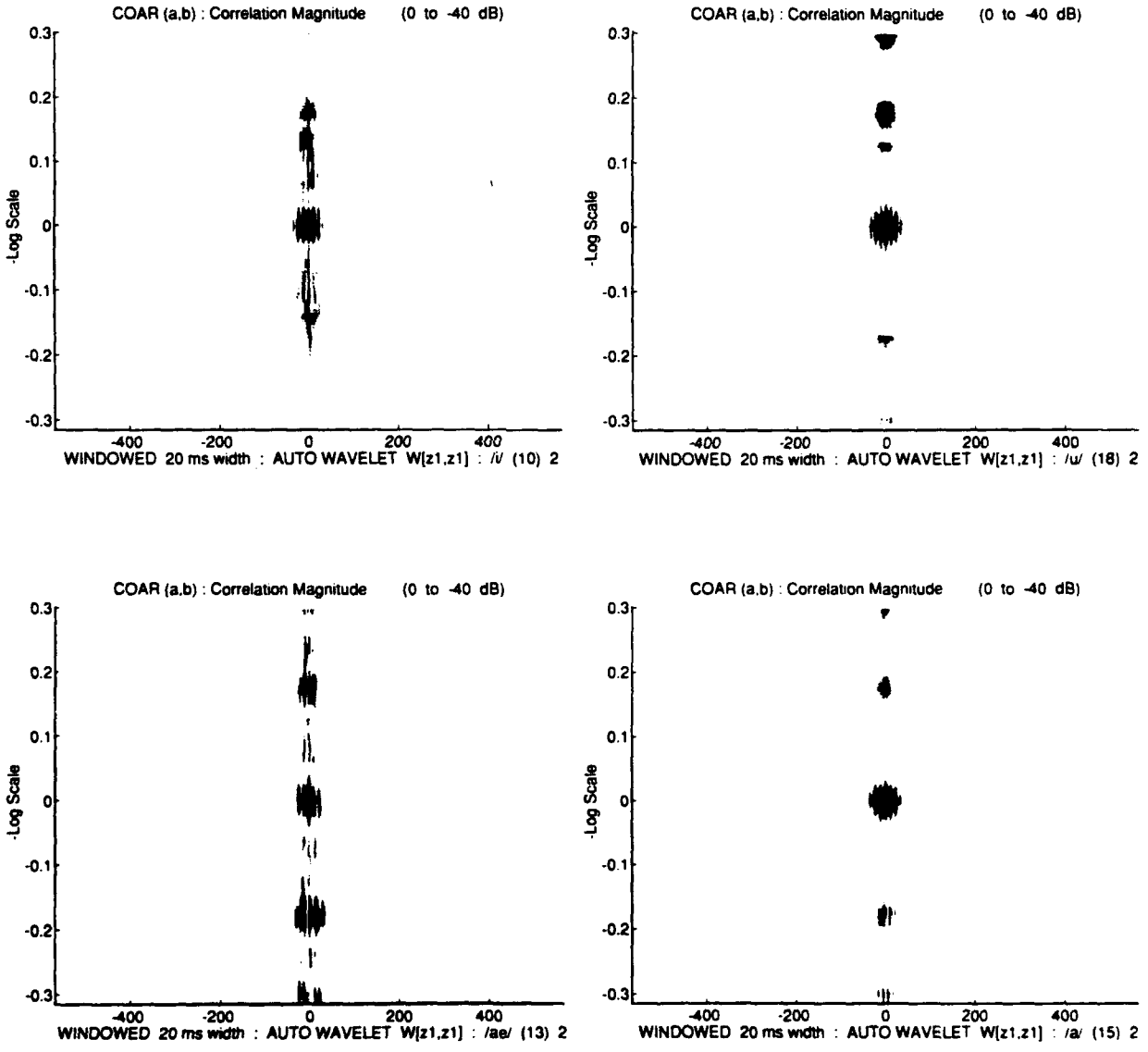


Figure 8.5 Auto-Ambiguity function:
[WINDOWED /V/, WINDOWED /V/]

Zoomed Time Auto-Ambiguity: [WINDOWED /V/, WINDOWED /V/]

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

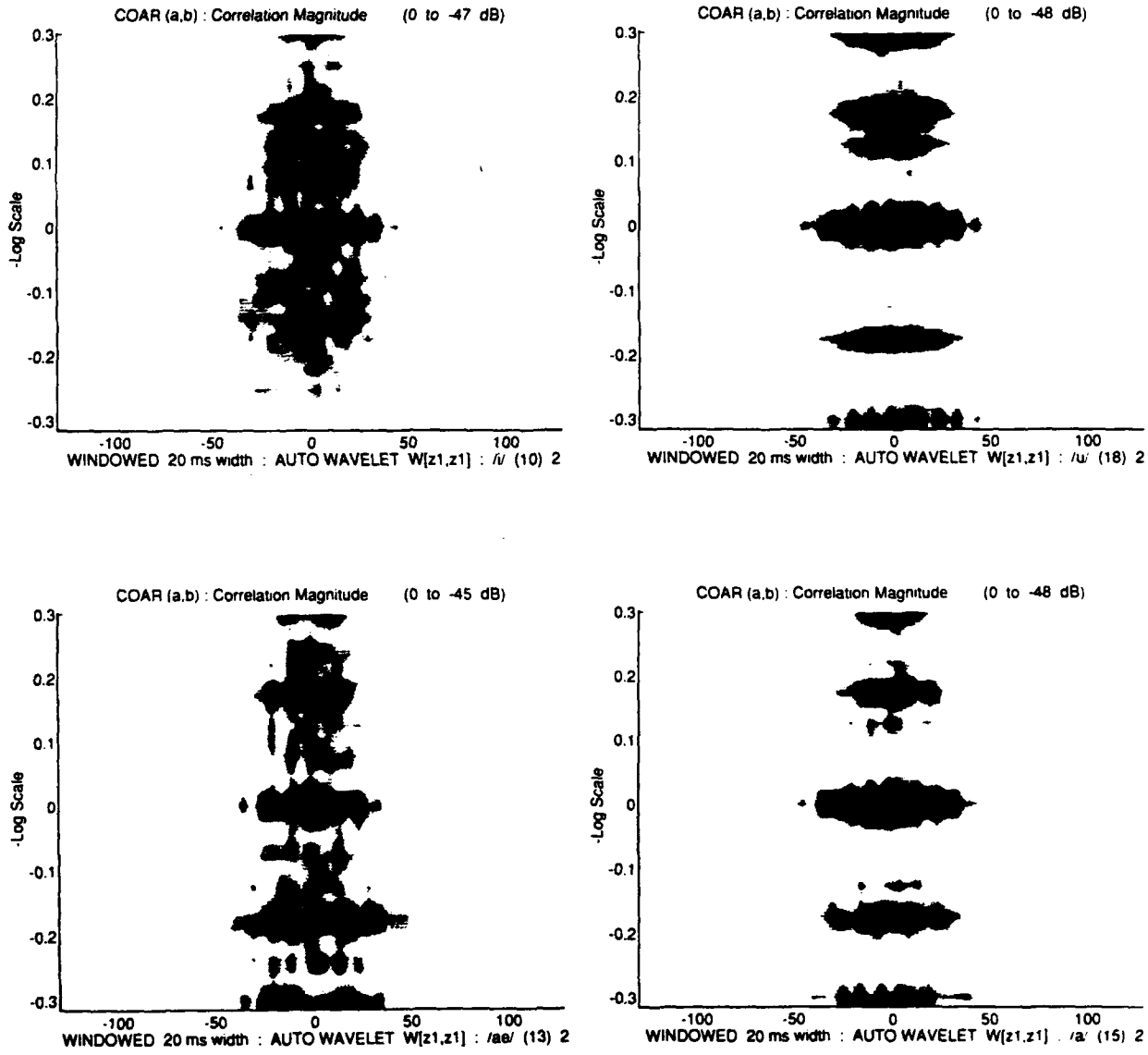


Figure 8.6 Zoomed Time Auto-Ambiguity:
[WINDOWED /V/, WINDOWED /V/]

In these plots, therefore, the vowels are contrasted by virtue of their articulatory differences, rather than by any aspect of their vocalization (glottal excitation).

The plots of Figures 8.5 and 8.6 additionally represent the time-frequency resolution of a vowel when used in the role of an analyzing wavelet. The relative compact-ness (in the time/scale plane) observed from the auto-ambiguity function is a measure of that signal's power-to-resolve in time and scale (Young 1993, pp. 175-180). In the case of the present vowels, for example, the diffuse structure of the auto-ambiguities generated from /i/ and /æ/ lie in contrast to the stark, centralized appearance of those from /ä/ and /u/. The difference implies that the latter vowels, /ä/ and /u/, offer superior scale-resolution, whenever they serve as an analyzing wavelet for a given **[CÔAR]**(a,b). This contrast may be a partial explanation for the apparent success of these vowels (particularly /u/) in generating distinct patterns of **[CÔAR]**(a,b) ridge trajectories.

8.3 Testing for the Null Case: COAR without the Coarticulation

The third empirical test on validation demonstrates the response of the **[CÔAR]**(a,b) distribution when an isolated vowel is crossed with (another repetition of) the same *isolated* vowel. In other words, what is the coarticulation channel when no consonants are spoken? This test documents a type of "null state" for the coarticulation model. The comparison differs from the previous auto-ambiguity function analysis, in that two *separate* utterances (and two distinct signals) are paired in the cross wavelet calculation.

In concept, the **[CÔAR]**(a,b) distribution associated with the null state is expected to be a mathematically trivial function, e.g., a lone spike located at the time/scale origin. Such a result would indicate that the coarticulation channel need only scale the input by exactly 1.0, and time-shift by exactly 0.0, in order to yield an effective reproduction of the *same* vowel at the channel output. It is known from examination of the auto-ambiguity functions, however, that this is not the case. An estimate of the **STV** channel which maps the same utterance into itself yields, at best, the auto-ambiguity function of that utterance.

In practice, therefore, it is expected that a test for the null case should yield a mathematically complex distribution, like the auto-ambiguity. *Un*like the auto-ambiguity, however, the null case introduces a *new* repetition of the same vowel in the model formulation (rather than a replica of the same vowel). A more appropriate comparison for a test of the null case is thus made in the context of a *complete* [/V/, C/V/C] utterance pair. The null state, in short, is most comparable to an ordinary coarticulation channel formulation, with the exception that another isolated /V/ plays the role of the effected utterance.

The following figures present this comparison. The first, Figure 8.7, depicts the windowed **[CÔAR]**(a,b) measured "normally" for the context n/−/n. In other words, Figure 8.7 contains the cross wavelet distributions calculated for the following coarticulation pairs: [/i/, "neen"], [/æ/, "nan"], [/ä/, "non"], and [/u/, "noon"]. Notice the presence of the familiar coarticulation landmarks attributable to the n/−/n context. Displaced ridge trajectories are visible from the pairs [/i/, "neen"] and [/u/, "noon"].

Channel Estimate: WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ n/V/n

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

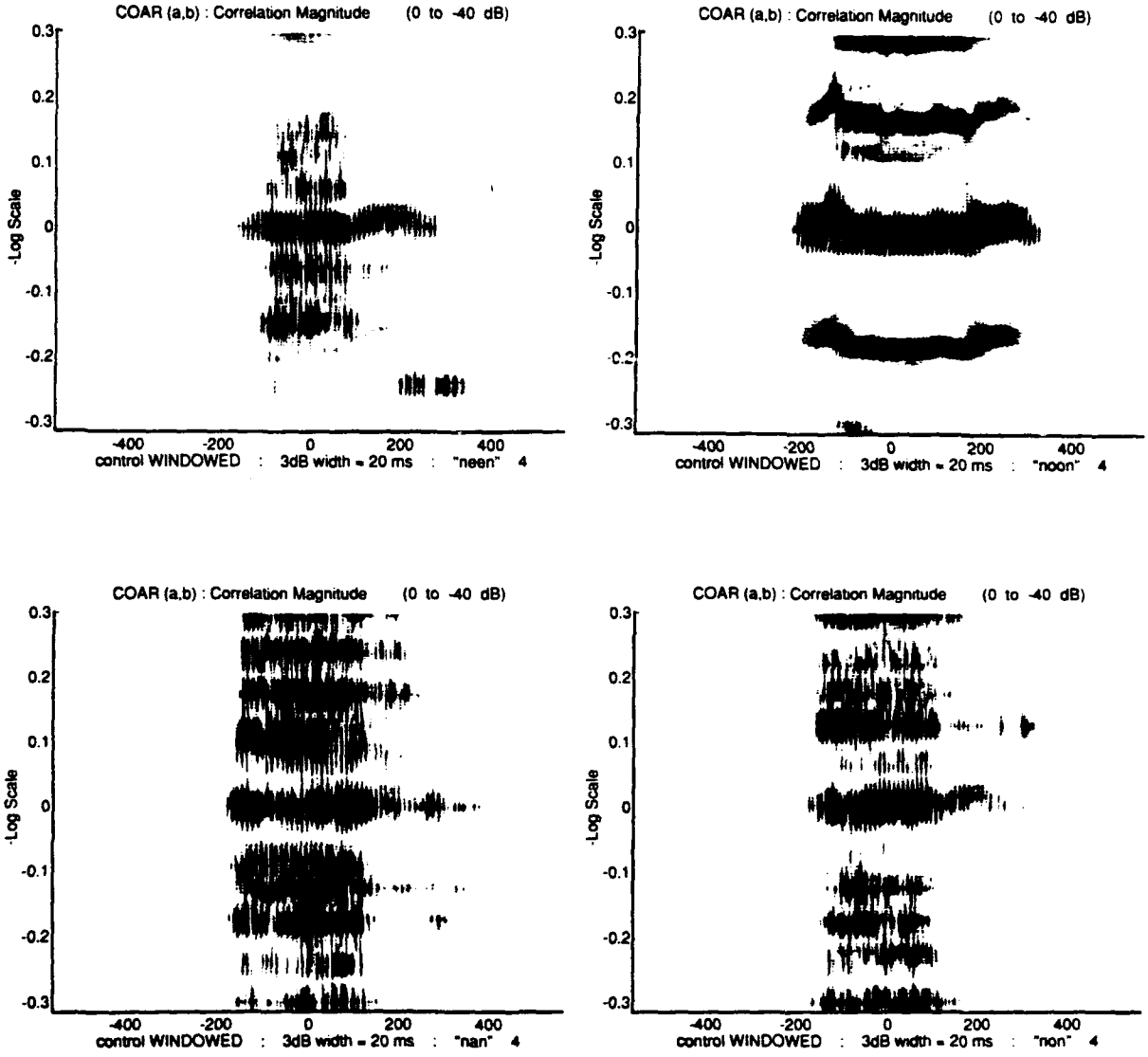


Figure 8.7 Channel Estimate:
WINDOWED /V/ ⇒ CÔAR(a,b) ⇒ n/V/n

Null State Channel: WINDOWED /V_1/ ⇒ CÔAR(a,b) ⇒ /V_2/

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

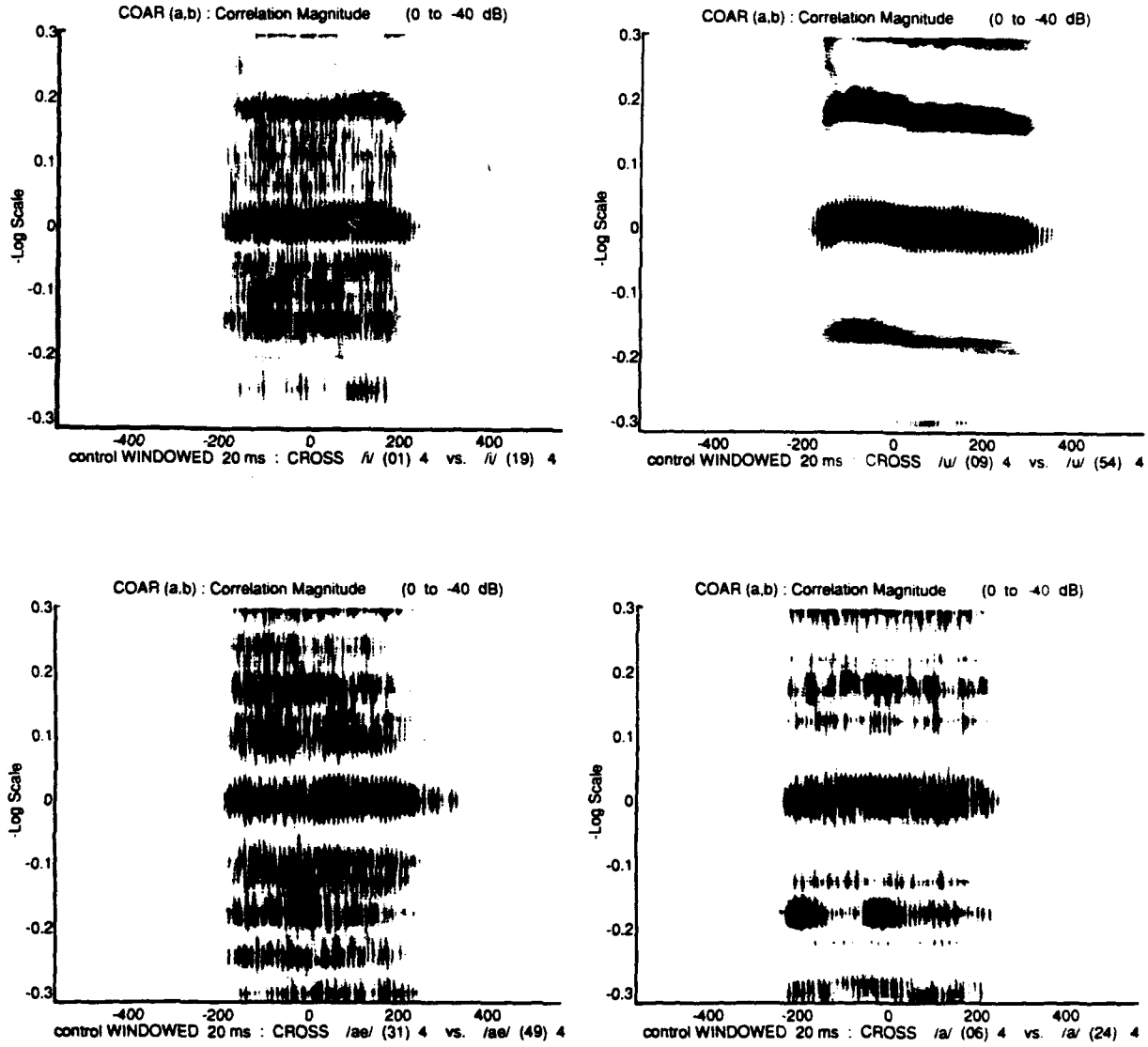


Figure 8.8 Null State Channel:
WINDOWED /V_1/ ⇒ CÔAR(a,b) ⇒ /V_2/

Ridge magnitude fluctuations, extending across 300 ms intervals, are exhibited in the other pairs, [/æ/, "nan"] and [/ä/, "non"].

Compare these plots with those of Figure 8.8, which illustrate the coarticulation channel estimates of the null state. The latter figure depicts the cross wavelet distributions formulated between one isolated vowel and another repetition of the same vowel. More specifically, Figure 8.8 shows the following cross wavelet pairs:

[WINDOWED $/i_1/$, $/i_2/$], [WINDOWED $/u_1/$, $/u_2/$],

[WINDOWED $/æ_1/$, $/æ_2/$], [WINDOWED $/ä_1/$, $/ä_2/$].

The subscripts (1,2) denote separate repetitions of the vowel. Although these null state estimates are hardly trivial in appearance, they do constitute a moderation of the familiar **[CÔAR]**(*a*,*b*) form. Notice, in the case of the vowel /u/, ridge displacements are present, but minimal. Other fluctuations in the magnitude of the correlation are visible to a limited extent.

Unfortunately, this null state comparison fails to yield any *conclusive* description on the behavior of the coarticulation channel. (Consider, however, that even an isolated vowel may contain some abrupt, consonant-like transients, e.g., the glottal stop at the onset of voicing. This would imply that a true null state for the coarticulation channel cannot be realized.) Nevertheless, the present comparison confirms that differences in the results do exist with respect to the absence or presence of some true consonantal context (in this example, n/−/n). Furthermore, this test documents the null state as a limiting case, and it provides a context from which all other (consonantal) cases may be evaluated.

8.4 Testing Overall Reproducibility of the Coarticulation Channel

The purpose of the final section on validation is to show that (using the normal configuration of the coarticulation channel) successive repetitions of the utterance pair yield **[CÔAR]**(*a*,*b*) distributions which are reasonably consistent from one repetition to the next. The comparisons which support this conclusion are stated simply. Each figure contains four different repetitions of the same [/V/, C/V/C] utterance pair. Within a figure, therefore, it is expected that the calculated coarticulation channel estimates will *not* vary significantly among repetitions.

Three such figures are presented. Each illustrates a different example coarticulation pair. *Between* figures, therefore, variations are expected.

1) Figure 8.9 contains the **[CÔAR]**(*a*,*b*) distributions calculated from four repetitions of the [/ä/, "dodd"] coarticulation pair.

2) Figure 8.10 contains the **[CÔAR]**(*a*,*b*) estimates from repetitions of the [/æ/, "gag"] pair.

3) Figure 8.11 contains the **[CÔAR]**(*a*,*b*) estimates from repetitions of the [/i/, "leel"] pair.

Notice the variability in the time base among these plots. Some variation in the plots' overall time-support arises from relative differences in the durations of their constituent C/V/C utterances. In other words, the duration of a C/V/C token is reflected directly in the time-support of the resulting **[CÔAR]**(*a*,*b*).

In effect, each figure is a *qualitative* measure of the within-group variability of the coarticulation channel. The measure of variation attributable to different phonetic

4 "dodd" Estimates: WINDOWED /ä/ ⇒ CÔAR(a,b) ⇒ d/ä/d

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

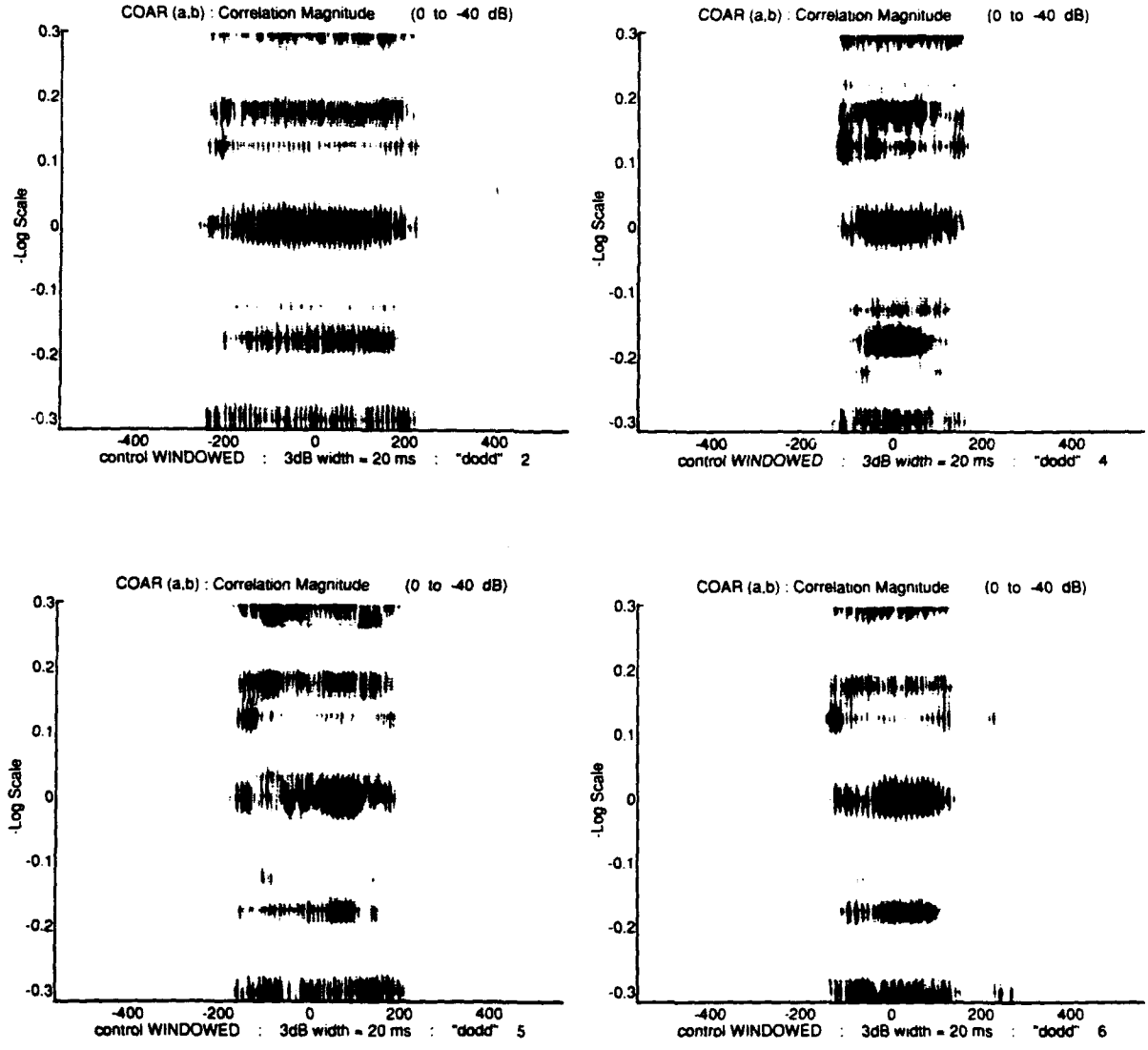


Figure 8.9 4 "dodd" Estimates:
WINDOWED /ä/ ⇒ CÔAR(a,b) ⇒ d/ä/d

4 "gag" Estimates: WINDOWED /æ/ ⇒ CÔAR(*a*,*b*) ⇒ g/æ/g

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

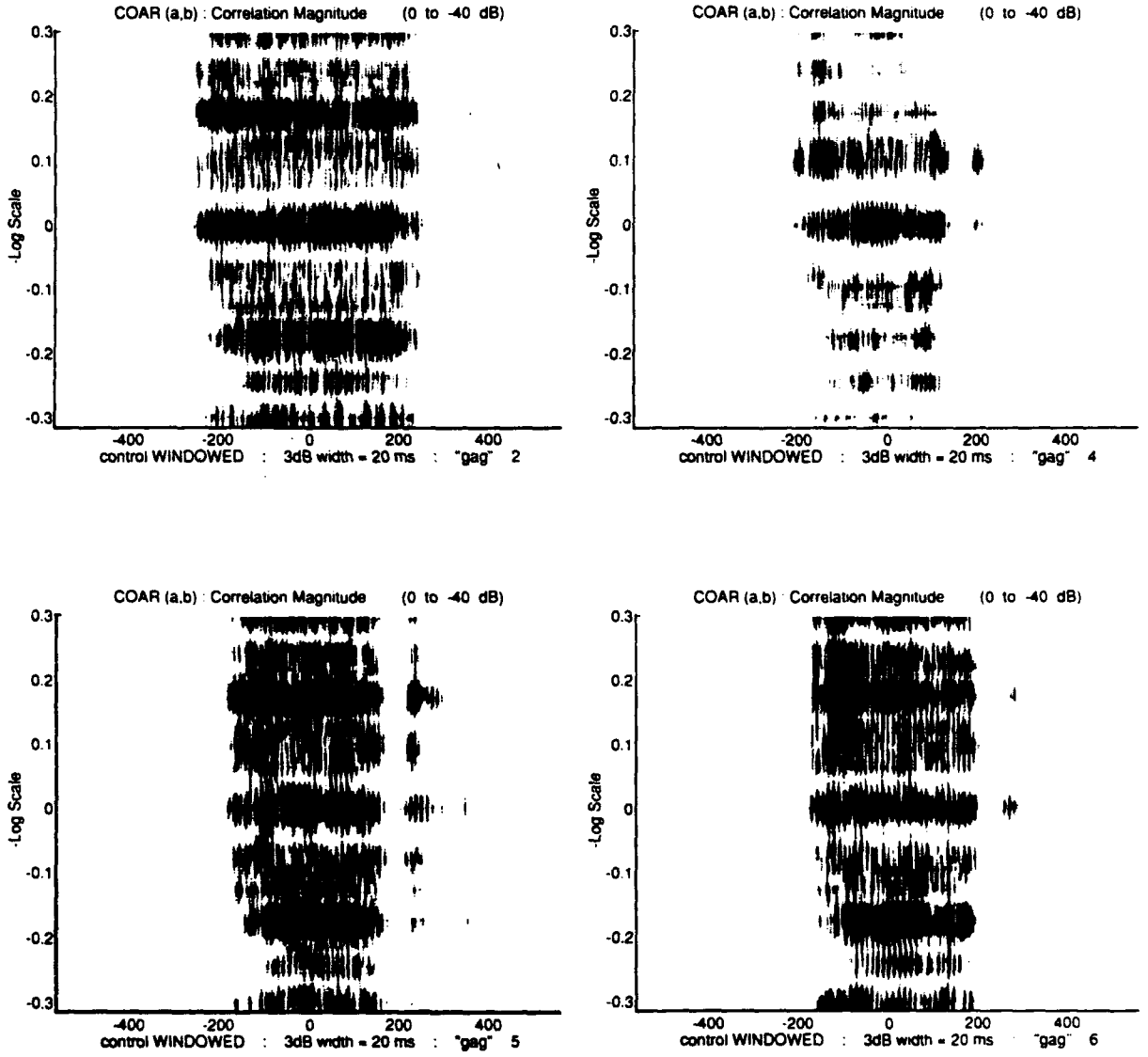


Figure 8.10 4 "gag" Estimates:
WINDOWED /æ/ ⇒ CÔAR(*a*,*b*) ⇒ g/æ/g

4 "leel" Estimates: WINDOWED /i/ ⇒ CÔAR(a,b) ⇒ l/i/l

(0 to −40 dB) vs. −Log Scale vs. Time-Shift (milliseconds)

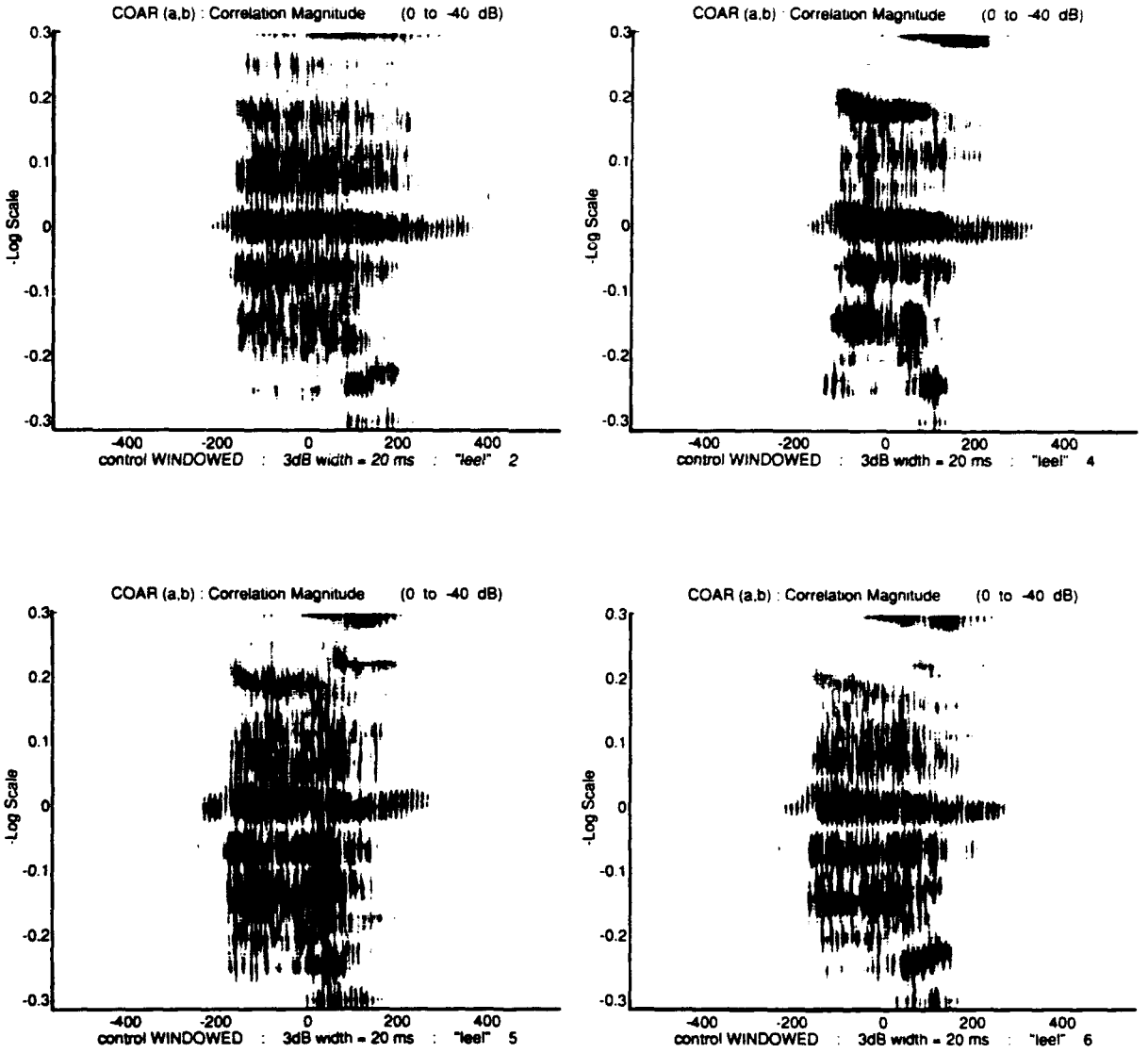


Figure 8.11 4 "leel" Estimates:
WINDOWED /i/ ⇒ CÔAR(a,b) ⇒ l/i/l

groups thus appears between different figures. Visual examination of these plots indicates consistency within a given figure and good reproducibility within the phonetic group. The between-figure variability appears to rival or, in some cases, exceed the variability within the group.

Due to the limitations of this visual assessment, no formal statistical conclusions can be drawn from it. For the sake of reference, however, this set of data results is a means for comparing a number of repeated evaluations of the model. Within the scope of these comparisons, it is concluded that individual instances of a particular [/V/, c/V/c] pair yield reasonably representative results, as manifested in the plots of these **[CÔAR]**(*a*,*b*) distributions.

Chapter 9

CONCLUSION

9.1 Conclusions Drawn from Theoretical Development of the Model

The proposed model for CVC coarticulation operates in the manner of an analysis-through-contrast. A control vowel which is free from the effects of coarticulation is contrasted with an effected version of that same vowel. The contrast is measured in the domain of the affine group. That is, differences between the two vowels are manifested in the scale-factor domain and time-shift interval domain. These differences are accounted for by the "coarticulation channel" and its characterization function: **COAR**(a,b).

The signal which most strongly characterizes the acoustic content of the vowel is the vocal tract noise-response function. Two such noise-response signals, therefore, are contrasted within the framework of the model. A measure of contrast between these signals is attained whenever an element of their similarity becomes displaced in either of the scale/shift dimensions. The model thus analyzes vowel differences by a method of "displaced commonality."

The mechanism for deriving such a contrast between two signals is the wavelet transform. In particular, the channel characterization function appears as the wavelet transform of the response signal from the CVC vowel, using the response signal from the isolated vowel as the analyzing wavelet.

Because the wavelet transform is invertible, it can be used for signal reconstruction. For this reason, the model's analysis-through-contrast may be interpreted equally well as a formula for re-creating the effected signal. The various (scaled and shifted) versions of the control signal serve as the ingredients for this re-creation.

From a theoretical standpoint, the appropriateness of this model as a tool for analyzing the coarticulation effect is very much invested in the tendency for that effect to assume the form of some *scaling* operation. There is little a priori evidence, however, that the CVC coarticulation effect is likely to assume this form.

9.2 Conclusions Drawn from the Theoretical Solution of the Model

In order for the model to be implemented in practice, its characterization function **COAR**(*a*,*b*) must be expressible in quantities that can be readily realized and measured. The **COAR**(*a*,*b*) function, defined in terms of two vocal tract noise-response functions, can be solved into an expression which depends on two *measurable* signals, namely, the voice output response signals.

This solution, however, utilizes the assumption that the conditions of voicing for the two vowels are uniform. Such an assumption can be reasonably satisfied when intensity level and pitch are carefully controlled. The consequences of these controls, however, may be to the detriment of an utterance's "natural" articulation. Just as the coarticulation effect is dependent on the *phonemic* context, some other aspects of articulation may be dependent on whatever *voicing* context is mandated by such controls.

With regards to the present study, it is not known what detrimental effects, if any, may have been introduced by deliberat)n the vowel's voicing.

9.3 Conclusions Drawn from the Experimental Study

The Morlet wavelet transform distributions calculated for the utterances of this study are reasonable representations of speech sounds. They can be interpreted appropriately as time-frequency analyzers, and they compare favorably with what is already known about these utterances from classical spectrograms.

The variable time-frequency resolution of the Morlet wavelet transform, however, distinguishes it from the spectrogram, and therein lies its strength. The superior time-resolution at high frequencies would be quite valuable in a speech analysis application specifically designed to measure temporal relationships. Judging from the isolated, single-syllable utterances analyzed in this study, it does not appear that the transform's inferior time-resolution at low frequencies would prove to be much of a drawback. Even in those applications where the times of voicing onset and offset must be pinpointed, there are usually many other cues (such as harmonics), which accompany the ridge fundamental. These other cues could be used for determining the onset of the voicing, in a frequency region where the higher time-resolution presides.

Likewise, it would appear that the good frequency resolution available from the Morlet wavelet transform at low frequencies would be quite advantageous for *measuring* fundamental pitch frequency. On the other hand, for the purposes of routine formant

frequency estimation, the greatly expanded *F1* harmonics could have a blurring effect on the location of that formant's average frequency value.

The study has generated a significant body of wavelet transform data on CVC articulations which was not available previously. The utterance set is a phonetically varied and balanced sampling of speech for a single subject.

The practical evaluation of the coarticulation channel has shown that the model does perform the function of an analyzer of CVC coarticulation effects. The description that it provides, however, appears to be effective for a limited class of CVC articulations. The role of the model as a *general* means of examining segmental coarticulation has not been definitively identified in this study. This is because the model's performance varies so greatly from one vowel to the next. Added to this are some basic questions concerning what is being represented (in an articulatory sense) by the **COAR**(a,b) ridge peak.

However, with respect to those classes of CVC utterances which do generate coherent ridge structures in the **[CÔAR]**(a,b) some interesting observations have been made. In some cases, these coherent structures are drawn from utterances whose spectrograms do *not* exhibit a wealth of coarticulation. The conclusions pertaining to either of these limited cases, however, bear a *general* significance in understanding the acoustic behavior of consonantal/vowel transitions. The coarticulation model has shown the presence of a vocalic acoustic structure very close to (and coincident with) the locus of the consonant. That is not to say that the consonant is composed exclusively from vowel-like elements. It does suggest, however, that, for the purposes of perception,

information on the identity of the vowel may be made available to the listener on a *continuous* basis throughout the transition.

What the results of this study have not shown is that the quality of the vowel perturbation is an invariant function of the phonetic (consonantal) context. The data does suggest, however, that the vowel identity, despite the effects of the perturbations, may still be perceptually recoverable at any point along the interval of transition.

With regards to the stated goal of reducing the dimensionality of the coarticulation problem, therefore, it is uncertain whether a significant reduction can be achieved without some phonetically invariant relationship in the **[CÔAR]**(a,b). In the case of the vowel /u/, it is possible that an abbreviated form of the **[CÔAR]**(a,b) matrix could be used to recreate the output (coarticulated) vowel, given only the input (isolated) vowel. The appropriate criteria for removing data from the **[CÔAR]**(a,b) distribution has not been studied, however, and they are not known. Such a reduced matrix, if found for one particular utterance pair, must also function for other instances of the pair. In such cases, a dimensionality reduction would then be realized, in the sense that each particular [/V/, CVC] combination would behave consistently in relation to the abbreviated **[CÔAR]**(a,b) form.

The final major conclusion drawn from this experimental study is a statement pertaining to the methods of calculation and the robustness of the model. The coarticulation channel estimates, **[CÔAR]**(a,b), are fairly reproducible from one repetition to the next. Such reproducibility establishes the validity of the model as a speech analysis tool and confirms the integrity of the calculations.

9.4 Discussion of the Conclusions

This thesis is structured in the manner of a theoretical proposal answered by a practical implementation of theory. Much of the analysis regarding the outcome of this study has been presented within the framework of:

1) how well the model has been substantiated by the generation of favorable experimental results,

and 2) how well the results have been substantiated in terms of what type of behavior can be anticipated, a priori, from the model.

This paradox derives from the novelty of the techniques. The proposed model for coarticulation does not constitute an evolution of many existing theories in the speech literature. Rather, it appears that no signal model for coarticulation has been proposed previously. Likewise, the wavelet transform techniques used for implementing this model are, as yet, unconventional methods for analyzing speech. As a result, some uncertainties regarding the statement of results (from either the theory or experiment) will undoubtedly arise. Indeed, some of the results suggest that the **[CÔAR]**(a,b) distribution may be responsive to articulatory effects *un*related to CVC coarticulation (as it is presently understood).

Nevertheless, based on the deductive constructs inherent in the model and the previous practical knowledge available on [illegible]terances, it is concluded that, in certain cases of evaluation, the model highlights some affirmative relationships in coarticulatory behavior. Furthermore, a data-base of controlled utterances, and a host of wavelet and

cross wavelet implementation algorithms have been set into place, in support of other further investigations in this area.

In particular, the aspects of this model's behavior which are the least understood might be best illuminated through a series of succinct, discriminating tests. Such tests will be well supported by the present work. The results of such tests could be incorporated by way of structural *modifications* to the original model. In turn, as the model behavior becomes more thoroughly understood, any *new* observations, pertaining specifically to articulatory processes, become more probable. Without such an investigative framework, it is less likely that any true departures from the available knowledge would result.

An example of such an investigation (though it would not constitute a modification *per se*) is a consideration of the *phase* component in the **COAR**(a,b) result. As the wavelet transform of a real signal taken with respect to another real signal, the **COAR**(a,b) is a distribution of real amplitude coefficients. Only the magnitude structure given by these coefficients was considered in the study. However, there may be a potential source of information carried by the [+1 vs. −1] phase of these coefficients, i.e., the (+/−) "sign" of each amplitude component. For the **COAR**(a,b) plots appearing in the Results chapter, perhaps a given ridge contains (+/−) phase *changes* which are in some way governed by the underlying articulations.

If nothing else, the present study emphasizes the complexity of articulatory processes and the richness of the acoustic signals generated from those processes. The outcome of this study, rather than casting doubt on alternative strategies of analysis, is

better taken as an impetus for finding yet more new ways of viewing speech production and improving the techniques for evaluating it.

9.5 Potential Applications and Future Work

A CVC coarticulation system (input, channel, output) has been formulated in wavelet-transform terms. The proposed model is capable of representing speech coarticulation in a time-varying fashion (i.e., in the form of a time-frequency distribution). This functional transformation, depicted in the **COAR**(a,b), is quantitative and reversible. If the behavior of the **COAR**(a,b) distribution is found to be consistent for a given class of vowel consonant combinations, therefore, the coarticulation model can benefit the following applications:

1) Using a pre-calculated **COAR**(a,b) as a "template" for coarticulation, natural-sounding synthetic vowels could be synthesized from their elementary, isolated counterparts.

2) By inverting the **COAR**(a,b), a naturally spoken vowel from context could be "reduced" into a form closer to that which is produced discretely. Such a reduced form would be more readily identified in a computer recognition scheme.

Appendix A

SELECTION OF THE ANALYSIS MOTHER WAVELET

The proposed coarticulation model is a system characterization of the behavior of a channel. This characterization is expressed in terms of the channel input and output signals. If, instead, the mere content of the signals themselves were of interest, then the analysis results would be subject to the following condition: The wavelet transforms of these input/output signals are subject to (they are a function of) the mother wavelet which is employed in their transformation. Indeed, the wavelet coefficient distribution for a given signal may change entirely from one mother wavelet function to the next; the wavelet transform is said to be "parameterized" by the mother wavelet.

The interest here, however, is in the relationship *between* the input and output signals of the speech-effect channel (Figure 4.2). This relationship is manifested in the estimate for the channel representation, $\mathbf{[C\hat{O}AR]}(a,b)$. Stated in equation [4.5], it appears as the "cross-wavelet" between the input and output. Therefore, there is no *analysis* mother wavelet which bears directly on the significance of the channel representation in $\mathbf{COAR}(a,b)$. That is, aside from some "smoothing" of the resolution window within the scale and shift-parameter plane, $\mathbf{COAR}(a,b)$ is *not* parameterized by an analysis mother wavelet (Young 1993, chapter 5, pp. 177-178).

On the other hand, the expressions used for calculating $\mathbf{[C\hat{O}AR]}(a,b)$ rely on the individual wavelet transforms of the input and output signals (equation [5.1]). They also employ some standard wavelet transforms of the measured speech signals, taken with

respect to some analyzing mother wavelet, $f(t)$. These transforms appear in equation [5.7].

Therefore, a step which is necessary for estimating the **COAR**(a,b) includes the calculation of some standard wavelet transforms and a selection of the analysis mother wavelet [$f(t)$] to be used in these transforms. But the only purpose in this regard is to calculate **COAR**(a,b); there is no intent to *characterize* how the estimate for **COAR**(a,b) responds specifically to the influence of the analysis mother wavelet.

A good choice for a mother wavelet in the analysis of any speech signal is a *non*-orthogonal one. The use of an *orthogonal* basis of wavelets (for the discrete wavelet-transform case) results in a scale-grid sampling which is too sparse for speech (Young 1993, pp. 51, 127). In other words, a speech signal contains many slight but distinct variations in frequency (such as formant transitions) which, when analyzed, correspond to scale values very near 1.0. Such a signal is therefore better suited to a *non*-orthogonal (sometimes known as "continuous") wavelet representation.

Another good choice for the mother wavelet when analyzing speech is one which yields a wavelet distribution that is immediately meaningful. Considering, still, the relative void of available wavelet data on real speech, it is necessary to compare one's own wavelet representation of a speech utterance to other *classical* representations of the same utterance. New wavelet data can only be interpreted on the basis of previous knowledge, so that a wavelet distribution which can be readily compared to classical representations is very much favored.

These considerations suggest the use of the Gaussian-windowed monochromatic pulse, otherwise known as the Morlet mother wavelet, $f_M(t)$ (Grossmann et al. 1989):

$$f_M(t) = e^{j\omega_0 t} \cdot e^{\left(-\frac{t^2}{2}\right)}$$

$$\omega_0 = 41.77 \text{ ms}^{-1}$$

$f_M(t)$ is a non-orthogonal wavelet which, by virtue of its "tonal" aspect, is particularly well-suited for analyzing speech resonances. The same complex-exponential kernel, $e^{j\omega t}$, appears (in a structurally different way) in all classical models of speech analysis. In fact, it can be shown that the wavelet transform taken with respect to this mother wavelet is equivalent to a *constant Q* (variable time-window) short-time power spectral analysis (Young 1993, p. 73). Figure A.1 plots the real part of $f_M(t)$:

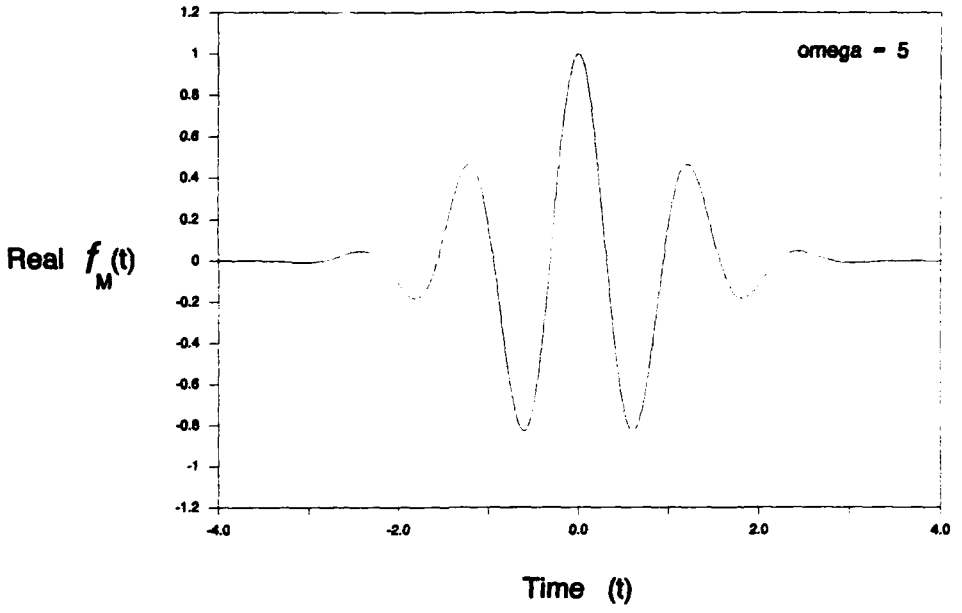


Figure A.1 **The Morlet Mother Wavelet $f_M(t)$**

The function $f_M(t)$ remains as the prototype mother wavelet used in some of the earliest wavelet research performed by Kronland-Martinet, Morlet, and Grossmann. Their study "Analysis of Sound Patterns Through Wavelet Transforms" included some samples of real speech (Kronland-Martinet et al. 1987).

Appendix B

INVERSION OF THE P_{SE} CHANNEL

The purpose of this appendix section is to define the inverse speech-effect channel and outline its associated estimate. The relationship between the forward and inverse forms of the $[\hat{P}_{SE}]$ is also derived. This inversion of the P_{SE} channel is pertinent to the applicability of the proposed model to problems in speech recognition (page 40).

The "forward" speech-effect channel P_{SE} transforms a control utterance $w1(t)$ into an effected utterance $w2(t)$, according to the effect-characterization estimated in $[\hat{P}_{SE}](a,b)$. As given in equation [4.3]:

$$w2(t) = \underset{P_{SE(a,b)}}{\mathbf{STV}} [w1(t)] = \int \frac{1}{a^2} \int P_{SE(a,b)} \frac{1}{\sqrt{|a|}} w1\left(\frac{t-b}{a}\right) db\ da$$

Let the *inverse* speech-effect channel be denoted P_{SE}^{-1}, and let it be specified by the following expression:

$$w1(t) = \underset{P_{SE}^{-1}(a,b)}{\mathbf{STV}} [w2(t)]$$

where, as before, $w1(t)$ and $w2(t)$ are the waveforms associated with the control and effected utterances, respectively. Figure B.1 illustrates the transfer associated with the P_{SE}^{-1} channel:

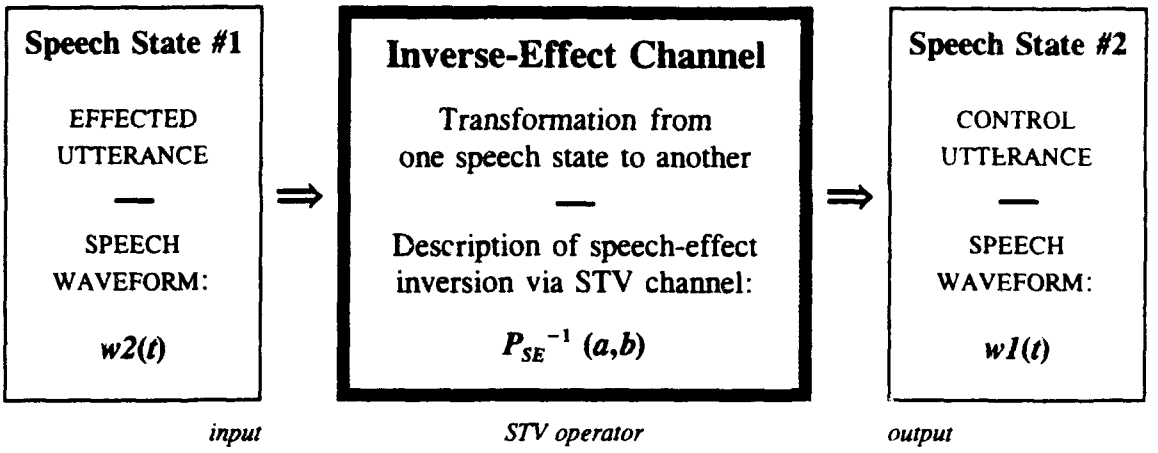


Figure B.1 **The Inverse Speech-Effect Waveform Channel**

Because P_{SE}^{-1} is so defined in terms of an **STV** channel characterization, the associated operation appears just as in equation [3.5]. Thus, $w2(t)$ is substituted for the input, and $P_{SE}^{-1}(a,b)$ is used as the $P(a,b)$ channel. The output $w1(t)$, therefore, is given by the following double integral:

$$\textbf{[B.1]} \qquad w1(t) = \underset{P_{SE}^{-1}(a,b)}{\textbf{STV}}\,[w2(t)] = \int \frac{1}{a^2} \int P_{SE}^{-1}{}_{(a,b)}\, \frac{1}{\sqrt{|a|}}\, w2\left(\frac{t-b}{a}\right) db\, da$$

Equation [B.1] specifies the **STV** operation for transforming a contextually effected utterance into an isolated control utterance.

The distribution $P_{SE}^{-1}(a,b)$ can be estimated just as any other $P(a,b)$ channel characterization. This estimate (equation [3.4]) appears as the wavelet transform of the output with respect to the input:

[B.2] $$[\hat{P}_{SE}^{-1}]\ (a,b) = \underset{w2(t)}{\mathbf{W}}\ w1(t)\ \ (a,b)$$

Notice how the estimate for the inverse P_{SE} contrasts with that of the forward P_{SE}. Recall from the previous section:

[4.2] $$[\hat{P}_{SE}]\ (a,b) = \underset{w1(t)}{\mathbf{W}}\ w2(t)\ \ (a,b)$$

The analysis which follows derives the specific relationship between the $[\hat{P}_{SE}^{-1}](a,b)$ and the $[\hat{P}_{SE}](a,b)$.

The wavelet transform used for estimating $P_{SE}^{-1}(a,b)$ in equation [B.2] is expanded according to its definition (equation [3.1]):

$$[\hat{P}_{SE}^{-1}]\ (a,b) = \underset{w2(t)}{\mathbf{W}}\ w1(t)\ \ (a,b) = \frac{1}{\sqrt{|a|}} \int w1(t)\ w2^{*}\left(\frac{t-b}{a}\right) dt$$

A change-of-variables is performed on the integral $\int dt$. Let:

$$t' \equiv \frac{t-b}{a} \qquad dt' = \frac{1}{a}\, dt$$

$$t = at' + b \qquad dt = a\, dt'$$

Then:

$$[\hat{P}_{SE}^{-1}]\ (a,b) = \frac{a}{\sqrt{|a|}} \int w1(at'+b)\ w2^{*}(t')\ dt'$$

$$= \sqrt{|a|} \int w2^{*}(t')\ w1(at'+b)\ dt'$$

$$= \left[\sqrt{|a|} \int w2(t')\ w1^{*}(at'+b)\ dt'\right]^{*}$$

Another substitution is made for the scale and shift parameters (a,b). Let:

$$\alpha \equiv \frac{1}{a} \qquad \beta \equiv -\frac{b}{a} \qquad \beta = -b\,\alpha$$

$$a = \frac{1}{\alpha} \qquad b = -\frac{\beta}{\alpha}$$

So that:

$$at' + b = \frac{t'}{\alpha} - \frac{\beta}{\alpha}$$

$$= \frac{t'-\beta}{\alpha}$$

Therefore, the estimate becomes:

$$[\hat{P}_{SE}^{-1}]\ (a,b) = \left[\frac{1}{\sqrt{|\alpha|}} \int w2(t')\ w1^{*}\left(\frac{t'-\beta}{\alpha}\right) dt'\right]^{*}$$

The expression in brackets is the wavelet transform of $w2(t)$ with respect to $w1(t)$:

$$\underset{w1(t')}{\mathbf{W}}\, w2(t') \quad (\alpha,\beta) \;=\; \frac{1}{\sqrt{|\alpha|}} \int w2(t')\; w1^{*}\left(\frac{t'-\beta}{\alpha}\right) dt'$$

This gives:

$$[\hat{P}_{SE}^{-1}]\ (a,b) \;=\; \left[\underset{w1(t')}{\mathbf{W}}\, w2(t') \quad (\alpha,\beta)\right]^{*}$$

The above wavelet transform constitutes the *forward* $[\hat{P}_{SE}]$ (equation [4.2]) stated in terms of the scale and shift parameters α,β:

$$[\hat{P}_{SE}^{-1}]\ (a,b) \;=\; \left[[\hat{P}_{SE}]\ (\alpha,\beta)\right]^{*}$$

Re-substituting the original scale and shift parameters (a,b) yields:

[B.3] $$[\hat{P}_{SE}^{-1}]\ (a,b) \;=\; [\hat{P}_{SE}]^{*}\left(\frac{1}{a}, -\frac{b}{a}\right)$$

Equation [B.3] thus shows how the forward and inverse $P_{SE}(a,b)$ channel estimates are related. Due to their mutual symmetry, the calculation of one estimate leads easily to the calculation of the other. As the wavelet transform of one waveform with respect to another waveform, the quantity $[\hat{P}_{SE}](a,b)$ describes a correlation between $w1(t)$ and $w2(t)$. The same is true for $[\hat{P}_{SE}^{-1}](a,b)$. When employed in the STV system operator, however, the opposing symmetry of these quantities differentiates between the "forward" and "inverse" directions of the speech-effect transformation.

REFERENCES

Basseville, M. (**1989**). "Detection of Abrupt Changes in Signal Processing," from *Wavelets: Time-Frequency Methods and Phase Space*, Combes, J.M., Grossmann, A., and Tchamitchian, Ph., editors. Berlin: Springer-Verlag, pp. 99-101.

Bendat, J.S. and Piersol, A.G. (**1986**). *Random Data*, New York: John Wiley and Sons.

Black, John W. (**1939**). "The Effect of the Consonant on the Vowel," *The Journal of the Acoustical Society of America* **10**, pp. 203-205.

Chistovich, L.A., Sheikin, R.L., and Lublinskaja, V.V. (**1979**). "Centres of Gravity and Spectral Peaks as the Determinants of Vowel Quality," from *Frontiers of Speech Communication Research*, Lindblom, B. and Öhman, S., editors. London: Academic Press, Inc., pp. 143-157.

Cole, R.A., Rudnicky, A.I., Zue, V.W., and Reddy, D.R. (**1980**). "Speech as Patterns on Paper," from *Perception and Production of Fluent Speech*, Cole, Ronald A., editor. Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp. 3-50.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J. (**1952**). "Some Experiments on the Perception of Synthetic Speech Sounds," *The Journal of the Acoustical Society of America* **24**(6), pp. 597-606.

Daubechies, Ingrid (**1990**). "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE Transactions on Information Theory* **36**(5), 961-1005.

DeFatta, D.J., Lucas, J.G., and Hodgkiss, W.S. (**1988**). *Digital Signal Processing: A System Design Approach*, New York: John Wiley and Sons.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (**1955**). "Acoustic Loci and Transitional Cues for Consonants," *The Journal of the Acoustical Society of America* **27**(4), pp. 769-773.

Fant, Gunnar (**1960**). *Acoustic Theory of Speech Production*, 's-Gravenhage: Mouton & Co.

Flanagan, James L. (**1972**). *Speech Analysis, Synthesis, and Perception*, Berlin: Springer-Verlag.

Flanagan, J.L. and Cherry, L. (**1969**). "Excitation of Vocal Tract Synthesizers," *The Journal of the Acoustical Society of America* **45**(3), pp. 764-769.

Fowler, Mark L. (**1991**). "Signal Detection Using Time-Frequency and Time-Scale Methods," Ph.D. Dissertation, Dept. of Electrical Engineering, The Pennsylvania State University, University Park, PA.

Grossmann, A., Kronland-Martinet, R., and Morlet, J. (**1989**). "Reading and Understanding Continuous Wavelet Transforms," from *Wavelets: Time-Frequency Methods and Phase Space*, Combes, J.M., Grossmann, A., and Tchamitchian, Ph., editors. Berlin: Springer-Verlag, pp. 2-20.

Grossmann, A. and Morlet, J. (**1984**). "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape," *SIAM J. Math. Anal.*, volume 15, pp. 723-736.

Harris, F.J. (**1978**). "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, **66**(1), January 1978.

House, A.S. and Fairbanks, G. (**1953**). "The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels," *The Journal of the Acoustical Society of America* **25**(1), pp. 105-113.

Jakobson, R., Fant, G., and Halle, M. (**1952**). "Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates," (Tech. Rep. 13) Cambridge, MA: MIT Acoustics Laboratory.

Kronland-Martinet, R., Morlet, J., and Grossmann, A. (**1987**). "Analysis of Sound Patterns Through Wavelet Transforms," *International Journal of Pattern Recognition and Artificial Intelligence* **1**(2), pp. 237-302.

Ladefoged, Peter (**1975**). *A Course in Phonetics*, New York: Harcourt Brace Jovanovich, Inc.

Liberman, A.M. and Mattingly, I. (**1985**). "The Motor Theory of Speech Revisited," *Cognition*, **21**, 1-36.

Liénard, J.S. and d'Alessandro, C. (**1989**). "Wavelets and Granular Analysis of Speech," from *Wavelets: Time-Frequency Methods and Phase Space*, Combes, J.M., Grossmann, A., and Tchamitchian, Ph., editors. Berlin: Springer-Verlag, pp. 158-163.

Lindblom, B. (**1963**). "Spectrographic Study of Vowel Reduction," *The Journal of the Acoustical Society of America* **35**(11), pp. 1773-1781.

Lindblom, B.E.F. and Studdert-Kennedy, M. (**1967**). "On the Role of Formant Transition in Vowel Recognition," *The Journal of the Acoustical Society of America* **42**(4), pp. 830-843.

Meyer, Yves (**1993**). *Wavelets: Algorithms and Applications*, Ryan, R.D., translator and editor. Philadelphia: Society for Industrial and Applied Mathematics.

Miller, R.L. (**1959**). "Nature of the Vocal Cord Wave," *The Journal of the Acoustical Society of America* **31**(6), pp. 667-677.

Monsen, R.B. and Engebretson, A.M. (**1977**). "Study of Variations in the Male and Female Glottal Wave," *The Journal of the Acoustical Society of America* **62**(4), pp. 981-993.

Neter, J., Wasserman, W., and Kutner, M.H. (**1990**). *Applied Linear Statistical Models*, Homewood, Illinois: Richard D. Irwin, Inc.

Nooteboom, S.G., Brokx, J.P.L., and de Rooij, J.J. (**1978**). "Contributions of Prosody to Speech Perception," from *Studies in the Perception of Language*, Levelt, W.J.M. and Flores d'Arcais, G.B., editors. Chichester: John Wiley and Sons, pp. 75-107.

Öhman, Sven E.G. (**1966**). "Coarticulation in VCV Utterances: Spectrographic Measurements," *The Journal of the Acoustical Society of America* **39**(1), pp. 151-168.

Öhman, Sven E.G. (**1967**). "Numerical Model of Coarticulation," *The Journal of the Acoustical Society of America* **41**(2), pp. 310-320.

Oppenheim, A.V. and Schafer, R.W. (**1975**). *Digital Signal Processing*, Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Peebles, Peyton Z., Jr. (**1987**). *Probability, Random Variables, and Random Signal Principles*, New York: McGraw-Hill Book Company.

Riley, Michael D. (**1989**). *Speech Time-Frequency Representations*, Boston: Kluwer Academic Publishers.

Rothenberg, Martin (**1973**). "A New Inverse-Filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing," *The Journal of the Acoustical Society of America* **53**(6), pp. 1632-1645.

Saito, S. and Nakata, K. (**1985**). *Fundamentals of Speech Signal Processing*, Tokyo: Academic Press, pp. 82-92.

Saracco, Ginette (**1987**). "Documentation du Programme Ondel: Logiciel de Decomposition de Signaux en Ondelettes". Laboratoire de Mecanique et d'Acoustique; Centre National de la Recherche Scientifique, France. Note technique number 5/87.

Schatz, Carol D. (**1954**). "The Role of Context in the Perception of Stops," *Language* **30**(1), pp. 47-56.

Searle, C.L., Jacobson, J.Z., and Kimberley, B.P. (**1980**). "Speech as Patterns in the 3-Space of Time and Frequency," from *Perception and Production of Fluent Speech*, Cole, Ronald A., editor. Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp. 73-102.

Sereno, J.A., Baum, S.R., Marean, G.C., and Lieberman, P. (**1987**). "Acoustic Analysis and Perceptual Data on Anticipatory Labial Coarticulation in Adults and Children," *The Journal of the Acoustical Society of America* **81**(2), pp. 512-519.

Steinhauer, K.M., Rekart, D.M., and Keaten, J. (**1992**). "Nasality in Modal Speech and Twang Qualities: Physiologic, Acoustic, and Perceptual Differences," *The Journal of the Acoustical Society of America* **92**(4), Part 2, 2pSP13, p. 2340.

Stevens, K.N. (**1972**). "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," from *Human Communication: A Unified View*, Denes, P.B. and David, E.E. Jr., editors. New York: McGraw-Hill, pp. 51-66.

Stevens, Kenneth N. (**1980**). "Property-Detecting Mechanisms and Eclectic Processors," from *Perception and Production of Fluent Speech*, Cole, Ronald A., editor. Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp. 103-112.

Stevens, K.N. and Halle M. (**1967**). "Remarks on Analysis by Synthesis and Distinctive Features," from *Models for the Perception of Speech and Visual Form*, Wathen-Dunn, W., editor. Cambridge, Massachusetts: MIT Press.

Stevens, K.N. and House, A.S. (**1963**). "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study," *Journal of Speech and Hearing Research* **6**(2), pp. 111-128.

Stevens, K.N., House, A.S., and Paul, A.P. (**1966**). "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," *The Journal of the Acoustical Society of America* **40**(1), pp. 123-132.

Strange, W., Verbrugge, R.R., Shankweiler, D.P., and Edman, T.R. (**1976**). "Consonant Environment Specifies Vowel Identity," *The Journal of the Acoustical Society of America* **60**(1), pp. 213-224.

Weiss, Lora G. (**1993**). "Wideband Inverse Scattering and Wideband Deconvolution of Acoustic Signals using Wavelet Transforms," Ph.D. Dissertation, Graduate Program in Acoustics, The Pennsylvania State University, University Park, PA.

Wornell, G.W. (**1990**). "A Karhunen-Loève-Like Expansion for $1/f$ Processes Via Wavelets," *IEEE Transactions in Information Theory*, volume 36, July, pp. 859-861.

Young, Randy Keith (**1991**). "Wideband Space-Time Processing and Wavelet Theory," Ph.D. Dissertation, Dept. of Electrical Engineering, The Pennsylvania State University, University Park, PA.

Young, Randy K. (**1993**). *Wavelet Theory and its Applications*, Boston: Kluwer Academic Publishers.